The Dissertation Committee for Adrian Rodríguez Riccelli certifies that this is the approved version of the following Dissertation:

The Subject Domain in Cabo-Verdean Creole: Combining variationist sociolinguistics and formal approaches

Committee:

Sandro Sessarego, Supervisor

Almeida Jacqueline Toribio

Chiyo Nishida

Lars Hinrichs

Nicolas Quint

The Subject Domain in Cabo-Verdean Creole:
Combining variationist sociolinguistics and formal approaches

by

Adrian Rodríguez Riccelli

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2019

Dedication

Dedico esta tesis a mis padres Armando y Mónica, a mi hermano Marcel, a mis hermanas Andrea, Emilia, y Claire, a mis abuelos Mauricio y Santiago, a mis abuelas Eulalia y Nora, a mis tías Claudia, Daniela, y Sandra, a mi querida Linda Concepción, y al ‘Big Gray’.

Acknowledgements

Completing this dissertation would have been impossible without the support, advice, direction, patience, loyalty, love, and friendship of my family, advisors, colleagues, teachers, friends, and my partner. I am forever indebted and grateful to you all and cannot thank you enough for helping me to achieve this; I will never forget it.

To begin with, I would like to thank the University of Texas Center for European Studies and Sally K. Dickson for their support with the Foreign Language Area Studies Fellowship, which came at a crucial early phase in the development of this project. I would also like to thank to the Graduate School and Dr. Mark J. T. Smith, Dean of the Graduate School, for their additional support in the form of the Dean's Prestigious Fellowship Supplement. Additional thanks go to the Department of Spanish & Portuguese, Dr. Jossiana Arroyo-Martínez, Dr. Jason Borge, Dr. Sergio Romero, and Dr. Almeida Jacqueline Toribio for, among many other things, their support with the Graduate Continuing Fellowship, which facilitated much of my travel for fieldwork and allowed me the time and resources to develop the project. Finally, I extend another thanks to Dr. Jossiana Arroyo-Martínez, as well as Dr. Sandro Sessarego, and Dr. Esther Raizen, College of Liberals Arts Senior Associate Dean for Research and Graduate Student Affairs, for their support in the final Summer and Fall semesters, without which completion of this project would have been impossible; thank you for believing in me!

The first and foremost of my gratitude must go to my advisor Dr. Sandro Sessarego, who saw me through the best and the most difficult moments of this journey, and who pushed me in all the right ways to grow as a scholar, a professional, and as a thinker. Sandro

arrived in the department at a time when the ideas for this project were at a nascent phase, when I was still pondering its feasibility, and when I was looking for someone to oversee my research. He gave the encouragement, guidance, and direction I needed to move forward. I thank you for inviting me to partake in collaborative research with you, for the countless professional opportunities, and for your advice and guidance in all matters academic. Sandro's loyalty and unwavering confidence in those in whom he sees potential drives him to extend his relentless work ethic to propping up others and ensuring their success. Particularly in the last year of this journey, when things often became very difficult, Sandro did not relent: he pushed me to strive forward and never let up, and believed in me even when things were not going our way. Beyond our work together, Sandro has been a dear friend and trustworthy confidant. I look forward to many years of collaboration and friendship, Sandro.

A second thanks is in order for Dr. Almeida Jacqueline Toribio, who saw and fostered my professional and intellectual development immediately upon my arrival here. Before I had even met her, Jacqueline's research was a major inspiration that stoked my curiosity in linguistics and the syntax of subjects when I still was just discovering the field. Her ability to explore a broad range of inter-disciplinary subjects and to transcend the theoretical and methodological constraints of the various subfields of linguistics is something I have always tried to emulate. It was an honor to take her classes and partake in many professional experiences with her, and to have benefitted from her sage guidance and advice. There were several moments when I pursed some unorthodox or seemingly unrelated coursework or line of inquiry, and though Jacqueline expressed skepticism and challenged me to justify some of these pursuits, in the end she believed in me and encouraged me to explore my curiosities to their logical end, and this is why many of the

innovations in this project are thanks to her. I will forever hold Jacqueline as a role model and inspiration, not just as a scholar, but as a whole person, since her unending genuine kindness is something we should all strive to nurture within ourselves.

A special thanks to Dr. Chiyo Nishida, not just for her inspirational work on clitics, but also for her invaluable advising, guidance, and collaboration. Chiyo oversaw several of my research projects at each important phase of the program. She provided me several valuable opportunities for collaboration that allowed me to develop my analytical skills in immensely useful ways. She also pointed me to canonical pieces of research at crucial points in my project that came to form a major part of the theoretical perspective adopted in this dissertation. Chiyo, you were also patient with me, giving me the room to develop my ideas and hone my writing. I cherish you as an advisor, colleague, and friend, and I am very much looking forward to our future collaboration and continued friendship.

Dr. Lars Hinrichs played another major role in my learning and growth over the entirety of the program at UT and throughout the development of this project. Lars in a excellent teacher, he showed me the ropes of using R and taught me so much about the fundamentals of sociolinguistics and language variation and change. He played no small part in inspiring me to pursue research in creole languages and morphosyntactic variation. He also lent essential guidance and support during the difficult stage of developing the quantitative analysis, much of the study design was informed directly by his advice. As a fun, charming, and relatable person, taking classes with Lars was always exciting and engaging, and I truly enjoyed every consultation with him. I look forward to the next NWAV, to hearing you on the Texas Standard, and to discussing a collaboration that I have been mulling over and that I will propose to you soon.

Dr. Nicolas Quint was another key figure in the realization of this project. I met Nicolas at ACBLPE 2016 in Praia when I was in Cabo Verde for my second summer of fieldwork. Without hesitation he extended his valuable friendship, guidance, and advice, and eagerly agreed to form part of the committee. Throughout my process of learning Kriolu and developing this project, he taught me about many key aspects of the language and clarified many doubts I had about particularl syntactic structure and constructions. His insights were invaluable for developing the historical and descriptive components of the project, and in particular for establishing the criteria for defining the envelope of variation. I am immensely grateful for his invitation to give a talk on my research at CNRS in 2018. During that stage of the project, I had developed the second-to-last iteration of the quantitative analysis, and the motivation to give that talk endowed the trajectory of the project with the momentum need to push me through to the end, and the feedback I received from Nicolas and colleagues informed directly the refinement and finalization of the research. I look forward to many future correspondences, chances for collaboration, and to eventually finding the opportunity to invite you to visit and give a talk at some future event that I can host.

Three collaborators and friends were absolutely essential to the completion of this project: Edmir Filomeno Freire Semedo, Clever, and Fáron Jesse D'jau Barbosa Peckham. All three of them contributed countless hours of editing, correcting, and adjusting interview transcriptions, as well as valuable comments and insights that allowed me to learn numerous lexical items, syntactic constructions, and other dynamics of Kriolu, all which helped me properly represent and analyze the data. Their tireless work is at the heart of this project, and needless to say, none of it would have been possible without them. Furthermore, in these three individuals, not only did I find valuable collaborators, but even

more valuable friends, and the opportunity to meet so many people and to truly get to know Santiago. I cannot thank Edmir and his mother Maria Filomeno Freire Lopes enough for receiving me, allowing me into their home, and housing me in the summer of 2016. It is difficult to describe how absolutely formative this experience was for me on both a personal and intellectual level, it was the truest and most genuine manifestation of *morabeza*. I learned so much Kriolu with Edmir, he introduced me to many informants, many of whom became friends, he partook in conducting interviews, took me on long hikes, and showed me so many of the things that are *terra terra. N kre pa-u sabi ma mi e iternamenti gratu. Abo e nha broda pa vida mo!* The same thanks and gratitude go to Fáron. In addition to his expert insights and skill in transcription editing, I am also grateful for the many fun times we had in Praia and Tarrafal, and cannot thank you enough for letting me stay at your place and for showing me so many interesting things about CV. I look forward to more collaboration and developing projects in the near future. Edmir, Clever, and Fáron, *anhos e realmenti kes spertu tras di kel projetu li. N ta spera ma nu ta teni munti anu di kolaborasôn i amizadi na nos frenti.* Thank you also to Jéssica Dias for her help and support with some of the transcriptions.

A very special thanks and my deepest gratitude goes two other dear friends and colleagues, Dr. David Giancaspro and Joshua Frank. David has been a most treasured friend since we began the Master's program at the University of Florida. I miss our days of pick-up basketball, racquetball, and raving about how underrated John Stockton and Dennis Rodman are. I treasure our discussions of politics, linguistics, and all other issues under the sun. David was endlessly patient and willing to spend many many hours with me through a grueling first year on the job market. Beyond that, his intellectual, professional, and personal advice has been among the most precious I received from anyone. I also hold

him as role model for his boundless kindheartedness, his unceasing loyalty as a friend, his intellectual prowess, and his well rounded balance of professional and personal life. Much of these same things could also be said of Joshua Frank. He was one the first friends I made when I got to UT and has remained a dear friend ever since. We shared great times in Austin and I also treasure our long conversations about all things that curious folks talk about: politics, linguistics, society, sports, culture. I am forever indebted to Josh for his help in reading, editing, and offering comments for the entirety of the dissertation. I am saddened that we wont be living in the same city anymore, but look forward to having you as friend, colleague, and collaborator for many years to come.

There are few other individuals from UT that I would like to thank as well. Thank you to Dr. Sergio Romero for your wonderful classes on sociolinguistics and your wise guidance and advice. Thank you to Dr. Vivian Flanzer for being an excellent Portuguese teacher, a dear friend, an excellent coordinator, and for giving me the opportunity to teach Portuguese. Thank you to my other excellent Portuguese teachers, Dr. Orlando Kelm and Dr. Daniela Miereles; Daniela, thank you for being such a great friend, I remember so fondly our dinners together. Thank you to Dr. Jocelly Guie Meiners and Dr. Delia Montesinos for their help and support when I taught Spanish for heritage language learners. I would also to thank Dr. Neil Kamil in whose class I developed much of the historical chapter and Dr. Keenan Pituch whose multinomial methods course directly informed the design of this study. Thank you to Dr. Arno Jacob Argueta, Dr. Natalie Rangel, and Dr. Brendan Regan for being excellent colleagues and wonderful friends.

I would be remiss to not mention some more friends in Cabo-Verde that were important along this journey. Thank you to Jacob Cullen, who I contacted before even arriving in Cabo Verde and who was essential in getting me set-up in and acquainted with

Praia, introducing me to many new friends, teaching me some of the basics of Kriolu when I first started learning, and for continuously being a great friend. Thank you as well to Zita Vieira, who was among the first friends I made in Praia. She introduced me to many friends and to her family, taught me a lot of Kriolu, and opened up the world of Cabo Verde for me. A special thanks goes to 'Jupa' Junhor Varela, a good friend whom I met in the third summer in Tarrafal. Jupa immediately made me feel at home and introduced me to all the wonderful people of *Kobôn Santxes*, the best neighbors anyone could ask for. *N ta gradisi kel amizadi i morabeza ki nhos da-n. Ó Jupa, txoma minis! So la!* Thank you also to Dr. Danae María Pérez for your friendship and valuable advice and support. Though I did not meet them in Cabo Verde, I would like to thank two formidable Cabo-Verdeanists, Dr. Marlyse Baptista and Dr. Bernardino Tavares for their correspondence and valuable insights. *També N-kre da nha gradisimentu di kurasôn pa tudu nhos Badiu i Djarfoguenzi ki partisipa na kel peskiza li i ki rasabê-n so ku morabeza i amizadi. Kauberdi e sabi pamodi nhos e sabi!*

I would like to thank Dr. Malte Rosemeyer for his correspondence and advice with respect to multinomial modeling. I would also like to thank Dr. Sonia Barnes for the skillful R workshop that she gave here at UT; it was incredibly helpful for begginers and all of us who attended have built on that early base of knowledge since then.

The true beginning of this journey goes back to my undergraduate studies at the University of New Haven, where several excellent professors led the way for me to become a critical thinker and budding scholar, and sparked my interest in history, society and eventually linguistics too. Thank you to Dr. Bradley Woodworth for being a skilled advisor and professor, for teaching me how to write, and for being a kind soul and nourishing my early intellectual curiosity. Thank you to Dr. Brett McCormick who was an excellent

professor and oversaw my senior thesis; when working on that project I encountered my first readings on linguistics. Thank you to Dr. Roberto Irizarry, whose Spanish for heritage learners classes was incredibly formative and important for me in my last year of undergraduate studies. It was in that class that I became certain I wanted to study Hispanic linguistics. I also treasure our discussions of Brazilian music and Latin American politics and culture. Thank you as well to Dr. Matthew Wranovix (I still recall your excellent course on Christopher Columbus!), Dr. Joshua Sandman, and Dr. Natalie Farringer.

I would also like to thank my advisors, professors, colleagues, and friends from the University of Florida; when I first arrived in Gainesville I knew nothing of linguistics, but you all fostered my growth and pushed me to excel. Thank you to my Master's advisor Dr. Ana de Prada Pérez for her abundant guidance and support and for showing me the ropes of the field. Dr. Jessi Aaron's course on language variation & change forever altered the way I think about language and view the entire natural world, I still think about it on a regular basis, remember it as among the best courses I have ever taken, and hope to model my own classes after it. Thank you to Dr. Jason Rothman, who pushed me to excel to the level of many of my more experienced colleagues and peers despite my initial unfamiliarity with the field. A special thanks to Dr. Eric Potsdam, among the best educators I have ever had the pleasure to learn from, his syntax seminars were some of the most challenging, but also some of the most fulfilling classes I have ever taken. I would also like to thank Dr. Gillian Lord Ward and Dr. David Pharies. A special thank you to Dr. Diego Pascual y Cabo for his ample advice and guidance, and who, along with Dr. Jennifer Cabrelli Amaro, Dr. Delano Lamy, and Dr. Tiffany Judy, were inspirational rolemodels for me when they were at the tail end of their graduate careers and when mine was just beggining. Thank you to Dr. David Giancaspro, Dr. Becky Halloran González, Dr. Osmer Balam, Dr. Dámaris

Mayans-Ramón, Dr. Ana María Díaz Collazo, Dr. David Vásquez Huratdo, Dr. David Miller, Dr. Sara Zahler, Dr. Katherine Honea, Dr. Anne Lingwall Odio, Dr. Carlos Andrés Bertolgio, Yanina Becco, Ana Solís, Sophia Durand, Elisabet Liminyana Vico, Carlos Enrique Ibarra, William Sheard, Meagan Day, Whitman Suárez, and Diana Pedraza, thank you all for being wonderful friends and inspirational colleagues.

I would also like to thank some folks here in Austin: Eldon Katz, Phill Janzen, Dan Nugent, Tommy Holden, and my dear brother Marcel; thank you all for being true and loyal friends, I might not have made it through this without you all.

I cannot express my endless love and thanks for my family. Thank you to my father, Dr. Armando Emilio Rodríguez, for cultivating in me a critical thinker and intellectual since childhood. Thank you for guiding me through life and for all your loyal and dedicated support, even when I would wander astray. Thank you for challenging me and teaching me to argue and debate. Thank you for your help with stats! Thank you for the many opportunities and for believing in me. You are such a core part of who I am. I love you deeply. Thank you to my mother Mónica Gilabert Riccelli. You have loved me, cared for me, guided me, nurtured me, I cannot possibly list or describe all the ways you have made me who I am. You saw me through thick and thin, and even at my most rebellious, you cared for unconditionally. You are the best mother, I love you forever, I love you for always. Finally, thank you my sweet and dear partner Linda Concepción García Sánchez. Your caring love, companionship, ceaseless support, and deep loyalty, saw me through this difficult journey. *Te amo con todo mi corazón.*

Abstract

The Subject Domain in Cabo-Verdean Creole: Combining variationist sociolinguistics and formal approaches

Adrian Rodríguez Riccelli Ph.D.

The University of Texas at Austin, 2019

Supervisor: Sandro Sessarego

This dissertation explores Subject Pronoun Expression (SPE) in Cabo-Verdean Creole (CVC), a Portuguese-based language spoken in the Republic of Cabo Verde. The CVC subject domain has at least three types of nominative anaphora: a subject clitic, a null subject, and a double-subject construction. This study is the first to examine the distribution of these subject categories by combining a quantitative methodology with formal syntactic theory, as well as insights from functionalist, usage-based, cognitive linguistic, and typological approaches. In so doing, it offers a new perspective on this issue that is intended to move the field past protracted theoretical debates over the morphosyntactic status and discursive functions of these grammatical elements. For instance, the formal category underlying subject clitics has been contested in CVC and cross-linguistically; some have claimed that they are independent pronouns that cliticize at the phonological level (Déprez 1994; De Cat 2005; Costa & Pratas 2013), others have identified them as inflectional affixes in the VP layer (DeGraff 1993; Baptista 1995; Culbertson 2010), while in language

typology they are analyzed as ‘person markers’ that can engage in local grammatical agreement or nonlocal anaphoric agreement (Bresnan & Mchombo 1987; Zribi-Hertz & Diagne 2002; Siewierska 2004; Creissels 2005; Kari 2017).

Sociolinguistic interviews and picture description narratives were collected from native speakers of CVC from the islands of Santiago and Maio. Sampled speech was transcribed prosodically (Chafe 1993; Du Bois *et al.* 1993; Torres Cacoullos & Travis 2019) in order to evaluate several aspects of discourse organization. Data were submitted to descriptive and inferential inspection in four analyses using R (R Core Team 2019): one was an exploratory test that served to delimit the variable context for SPE in CVC, the second involved a fixed-effects multinomial logistic regression, and the third and fourth were based on mixed-effects binomial logistic regressions.

Results revealed highly significant effects for linguistic structural priming: double-subject and singleton tonic pronouns primed subsequent double-subjects, while null subjects primed additional null subjects. Lexical Determiner Phrase (DP) antecedents that were semantically referentially deficient (i.e. they bore inanimate, indefinite, or nonspecific reference) also promoted anaphoric zeros. These results lend partial support to the claims regarding the semantic properties of strong pronominals proposed under the Typology of Structural Deficiency (Cardinaletti & Starke 1994, 1999), and suggest that, as in Brazilian Portuguese, there is an “avoid referentially deficient pronoun” constraint (Kato & Duarte 2003, 2005; Duarte & Soares da Silva 2016) that is probabilistically active in CVC.

The zero-to-zero priming effect and the favoring effect from referentially deficient lexical DPs were only active at short anaphoric distances, and were promoted when adjacent intonational units were prosodically linked or simultaneously prosodically and

syntactically linked (Torres Cacoullos & Travis 2019). The priming effect for double-subjects obtained at longer anaphoric distances; they are promoted when their antecedent is in a non-adjacent clause.

Results suggest that double-subjects function as switch-reference devices, can establish contrastive focus, and reintroduce old discourse referents. These are much the same functional and discursive values that singleton tonic pronouns have cross-linguistically (Givón 1976; 2001[1984]; 2017). The realization of zero subjects is mostly contingent on antecedent accessibility (Givón 1976; 2017, Ariel 1990), but is also modulated by the aforementioned "avoid referentially deficient pronoun" constraint.

Inferring from the results for zero and double-subjects, it appears that CVC subject clitics are 'ambiguous person agreement markers' (Bresnan & Mchombo 1987; Siewierska 2004): like independent pronouns, they engage in nonlocal anaphoric agreement, but like inflectional affixes, they also engage in local grammatical agreement. This in-between morphosyntactic status is related to the infinitival origin of CVC verbs (Quint 2008b): the absence of bound person-number inflection is likely to have initiated grammaticalization on tonic pronouns, causing them to be eroded into subject clitics, and eventually become ambiguous person agreement markers, which are probabilistically dropped according to the properties of their controllers and the dynamics of antecedent accessibility. In line with Wratil's (2011) 'Null Subject Cycle', it could be argued that CVC subject clitics are grammatical elements that have stagnated at an early stage of a grammaticalization cline, which entails the transformation of independent pronouns into clitics, and then eventually into bound affixes.

Table of Contents

List of Tables

List of Figures

List of Illustrations

List of abbreviations used in this dissertation

SPE	-	Subject Pronoun Expression
CVC	-	Cabo-Verdean Creole
SC	-	subject clitic
X2SBJ	-	double-subject construction
Ø	-	null subject/anaphoric zero
GG	-	Generative Grammar
NSP	-	The Null Subject Parameter
NSL	-	Null Subject Language
UGC	-	The Upper Guinea Coast
UGCP	-	Upper Guinea Coast Creole Portuguese
BP	-	Brazilian Portuguese
EP	-	European Portuguese
AP	-	Angolan Portuguese
EP	-	European Spanish
AS	-	Argentine Spanish
PRS	-	Puerto Rican Spanish
DS	-	Dominican Spanish
LOD	-	log odds
MLR	-	multinomial logistic regression

Chapter 1: Introduction.

The goal of this dissertation is to explore **Subject Pronoun Expression** in Cabo-Verdean Creole, a Lusophone-based language spoken in the archipelagic Republic of Cabo Verde adjacent Senegambia and the Upper Guinea Coast regions of northwest Africa. Cabo-Verdean Creole developed an inventory of nominative anaphora that distinguishes its subject domain from many other Ibero-Romance-based language contact vernaculars that developed throughout the Atlantic basin over the course of the colonial era: it has a **disjunctive tonic-atonic opposition** among its overt subject pronouns.

This study is the first to attempt to explore Subject Pronoun Expression (SPE) or nominative anaphora resolution in Cabo-Verdean Creole (CVC) by combining a **quantitative** or **variationist sociolinguistic** methodology with theoretical constructs from **formal theoretical syntax** and insights from **functionalist**, **usage-based, 'cognitive linguistic',** and **typological** approaches. As part-and-parcel of this methodological approach, language-external social factors that might exert some effect on variable SPE are also considered. Another goal of this study is to situate CVC typologically among post-colonial Ibero-Romance-based language contact vernaculars and other African and Romance languages that have atonic subject pronouns.

1.1 Introduction to the major themes in this dissertation.

As far as archaeological evidence and historical records have shown, the Cabo-Verdean archipelago was not permanently settled until the mid-15th century when *Ribeira Grande* (today *Cidade Velha*) became the first permanent European-founded settlement in the tropics (Lobban 1995:10,16). Since then, the nation of Cabo Verde has found itself at a point of intersection in the Atlantic world.

It is a society that emerged from the imperial expansionist fervor of the post-*'reconquista'* period when Iberian maritime powers pushed outwards into the Atlantic; settlers scrambled to colonize the Atlantic islands and merchants and navigators sailed down the West African coast seeking to trade in consumer goods and enslaved people. As a result, Cabo Verde was the first society founded amidst the volatility of the earliest phase of the European ultramarine colonial era and the violent demographic trends of its slavery-dependent economy. Even before the opening of the westward routes of the trans-Atlantic Slave Trade, a new people, society, culture, and language had already emerged from the encounter of Africans and Europeans on Santiago, the largest, most populous, and longest inhabited island of the Cabo-Verdean archipelago. Since the expansion of the Triangle Trade to American shores, the historical trajectory of the islands became inextricably linked with settlements to its west as well (O'Malley & Borucki 2017).

The geographic, historical, and cultural positioning of Cabo-Verde at the nexus of the three major spheres of the Atlantic world is manifested in the native language of Cabo-Verdeans, CVC. Perhaps as a reflection of this manifestation, the structure of CVC implicates overlapping lines of inquiry that have been concerned with exploring the morphosyntactic properties of languages from all three continents. Of all the features of this language that might serve as an example of this intersectionality, SPE is perhaps the best case-in-point, since, as we will see, CVC has syntactic reflexes in its subject domain that can be found in languages from each sector of the Atlantic world. It is for this reason that the study of SPE in CVC is an essential piece of the puzzle in the project to understand the consequences of language contact, variation, change, and new language formation in the vernaculars of the post-colonial Atlantic world, and thus why this topic is the focus of this study.

The CVC subject domain developed two sets of overt subject pronominals: a tonic set and an atonic one. This distinguishes CVC from American varieties of Portuguese and Spanish that were also shaped by language contact in the colonial era, although bipartite subject pronoun inventories with tonic-atonic oppositions can be found in other Portuguese-based creoles, elsewhere in Romance, and in many African languages. Depending on the theoretical perspective

of the researcher wishing to classify them, and the classification conventions that have historically been used for a given language or language family, atonic subject pronouns are sometimes labelled **subject clitics**, **person agreement markers**, or **anaphoric subject markers**, among other terms (Siewierska 2004:8,121-127; Kari 2017). Following Baptista (1995, 2002), Pratas (2004), Costa & Pratas (2013), and classification conventions in the Romance tradition, I will refer to CVC atonic subject pronouns/markers as **subject clitics** (SCs henceforth) (1).

(1) ***N***=*txoma* *Jas*
1.SG.CL=call J.
'I am called J.'

The counterpart to SCs is a set of **tonic** subject pronouns. For the tonic set of pronominals, a distinction is sometimes drawn between *a-* initial disyllabic **strong** forms and monosyllabic **weak** ones[1] (2). In CVC, with all verbs except individual-level copula *e* 'to be', tonic pronouns almost always co-occur with SCs in double-subject constructions (3).

(2) ***(A)mi*** *e* *berdianu*
1.SG.STR COP cabo-verdean
'I am Cabo-Verdean.'

[1] Some classifications take the strong-weak distinction among tonic subject pronominals to represent universally underlying morphological categories associated with a set of syntactic, semantic, and prosodic properties (Cardinaletti & Starke 1994, 1996, 1999; Baptista 2002; Pratas 2004), while others treat tonic forms as pertaining to single underlying class with some distributional and pragmatic tendencies noted for the disyllabic forms (Veiga 1996:177; Quint 2000a, 2000b; Lang 2012). See Section 4.1 for further discussion.

(3) ***(A)nos*** ***nu***=*ta* *fla ‘ag’,* ***(A)es*** ***es***=*ta* *fla ‘agu’*

1.SG.STR 1.SG.CL=TMA say ‘water’ 3.SG.STR 3.SG.CL=TMA say ‘water’

‘We say “ag”, they say “agu”’

As in many vernaculars shaped by the forces of language contact, CVC did not retain the suffixal inflectional morphology marking person-number subject-verb agreement on finite verbs that was found in its superstrate (Pratas 2004:43). Quint (2008b) argues that the reason for the absence of person-number inflection is because the Late Medieval/Early Classical Portuguese infinitive is the origin of the CVC bare verb.

The disjunctive CVC subject pronoun inventory reflects the bipartite subject pronoun systems of its major substrate languages, Wolof and Mandinka, which also have this tonic-atonic opposition, though CVC SCs do not undergo the complex transformations of Wolof subject markers. Other than relatively infrequent tonic subject pronouns, SCs are the only overt instantiation of nominative person-number marking in CVC and are by far the most common. Crucially, like in all varieties of Portuguese, the same set of third person overt pronominals that bear coreference with animate, definite, and specific discourse referents, can also resume inanimate, indefinite, and nonspecific antecedents; in the case of CVC, third person SCs usually carry out this function. The distribution of anaphoric expressions like (1), (2), and (3) and the dynamics of their anaphoric relationship to discourse antecedents are a matter for empirical confirmation and one of the major themes considered in this investigation.

The analysis that one adopts for the morphosyntactic status and underlying clausal position of SCs relative to the subject position will be determinative in how one views the status of null subjects. Indeed, the relationship between SCs and the availability of null subjects has been a point of contention in previous studies of the CVC subject domain (Baptista 1995, 2002 vs. Pratas 2004; Costa & Pratas 2008, 2013). These arguments generally proceed as follows: if SCs are bound inflectional affixes, then absent a lexical DP or double-subject, the subject position must contain

some empty category, whereas if SCs are anaphoric pronominals in the subject position, then any doubling from a lexical DP or tonic subject pronoun would suggest that these stressed forms are in a left-peripheral position. Yet, regardless of the underlying clausal position of SCs, and despite claims to the contrary (Costa & Pratas 2008, 2013), null subjects have been attested even in utterances with no overt subject element whatsoever (Baptista 2002:258-260). I too observe this use of anaphoric zero subjects (4) in the corpus of spoken CVC analyzed in the current study.

(4) *E*$_{i}$*=kai na lagoua, i ali, dja* ***Ø***$_{i}$ *labanta*
3.SG$_{i}$=fall in pond CONJ here TMA 3.SG$_{i}$ rise
'He fell in the pond, and here, now [he] rises.'

Another goal of this study is to ascertain the distribution of each of the subject elements in (1) – (4). However, as we will see, (2) rarely occurs in CVC except for with the verb *e* 'to be'. Therefore, the primary objective of the present study is to explore the distribution of (1), (3), and (4), and to develop probabilistic analyses to explore the combination of independent predictor factors that can be found to correlate with the realization of each.

In Generative Grammar (GG henceforth) approaches[2], as well as in the usage-based/functionalist/'cognitive linguistic'/typological tradition, the morphosyntactic status and/or hierarchical position of SCs has been contested; these contestations have hinged on similar conceptual constructs in both theoretical paradigms (Siewierska 2004:121). What is more, disagreements over the formal category and syntactic position of SCs have emerged repeatedly among scholars of many different languages.

One analysis attributes SCs the status of a **phonological clitic** or **independent anaphoric pronoun**. In other words, SCs are taken to be genuine subject pronouns that occupy a canonical

[2] In this dissertation, I will use the terms 'GG' or 'formal[ist]' approaches to describe the theoretical and methodological tradition associated with transformational grammar, the Minimalist Program, and frameworks such as Principles & Parameters.

subject position and only undergo cliticization or attachment to a host at the phonological level of processing and production (cf. Kayne 1975; Brandi & Cordin 1989; Déprez 1994; Pratas 2004; De Cat 2005; *inter alia*). The competing analysis identifies SCs as **bound inflectional affixes** or **person agreement markers** that mark person-number subject-verb agreement and are structurally attached to or dependent on their host, or are syntactic heads contained within the INFL layer (Givón 1976, 1983a,b,c,d, 2001[1984], 2017; Rizzi 1986a; Bresnan & Mchombo 1987; Brandi & Cordin 1989; DeGraff 1993; Baptista 1995, 2002; Siewierska 2004; Culbertson 2010; Kari 2017 *inter alia*). A major theoretical model from the GG tradition takes a slightly different approach: in the Typology of Structural Deficiency (Cardinaletti & Starke 1994, 1996, 1999), three universally underlying classes of pronominal[3] are identified: **strong**, **weak**, and **clitic**. Each class is associated with a set of morphosyntactic, prosodic, and semantic-referential features.

In light of the protracted nature of these debates, and the inability to conclusively identify the pronominal or affixal status of SCs using the classical 'tests' for distinguishing these elements (cf. Kayne 1975; Klavans 1982; Zwicky & Pullum 1983; *inter alia*), another objective of the present study is to take a first step towards breaking the theoretical stand-off over the status of SCs in CVC and cross-linguistically. Instead of applying syntactic tests to SCs themselves, I will examine the probabilistic distribution of the other two nominative anaphora that 'compete' with SCs for the 'subject space' in CVC, double-subjects[4] (3) (X2SBJs henceforth) and anaphoric zero (4). By taking SCs to be the 'baseline' form of overtly marking person-number (regardless of their pronominal vs. affixal status) and comparing (3) and (4) against them, one might be able to infer from the properties of the latter two, something about the status of the former.

As was noted above, the analysis that one ultimately attributes to SCs will also determine the conclusions that one draws with respect to the availability of null subjects or anaphoric zero subjects (Ø henceforth). The study of Ø subjects in the GG tradition has centered around the Null

3 The "typology of structural deficiency" is extended to non-pronominal elements as well, such as adjectives and adverbs, among others.

4 Henceforth, the abbreviation X2SBJs will be used to refer specifically to the tonic pronoun + SC type, as opposed to the lexical DP + SC type of double-subject.

Subject Parameter (NSP henceforth), a typological model to describe the distribution of Ø subjects and related morphosyntactic properties in the world's languages, and the underlying Principles & Parameters of the human language faculty that constrain that distribution. Under most instantiations of the NSP, there is a phonetically vacuous morpheme underlying [+argumental, +referential] Ø subjects known as *pro*, which requires **identification** to be licit (Chomsky 1981; Rizzi 1986b; Camacho 2013:68). For many Null Subject Languages there is a correlation between the availability of *pro* and the presence of 'rich' inflectional morphology marking subject-verb agreement with distinct inflections for each person-number instantiation; identification of *pro* is often (though not always) assumed to be carried out by this 'rich' inflection[5] (Cole 2009; Camacho 2013:112-115,152-152). Further, the widespread availability of *pro* in Null Subject Languages has long been assumed to correlate with a **cluster of morphosyntactic properties** under the NSP (see Section 4.5.1) (Perlmutter 1971; Chomsky & Lasnik 1977; Chosmky 1981; Gilligan 1987; Camacho 2008, 2013:39-66, 2016:28-30). Languages with 'rich' inflection, high rates of *pro* relative to overt subject pronouns, and which display the relevant cluster of properties associated with a positive status under the NSP, came to be referred to as **consistent Null Subject Languages**.

Later, the range of mechanisms capable of identifying Ø under the NSP was expanded, allowing for identification from discourse topics in discourse-oriented or topic-prominent languages (Huang 1984). Recently, Camacho (2011, 2013:112-115,146-173) has shown that topic-identified-*pro* may also be active even in consistent Null Subject Languages, allowing for antecedent-anaphor coreference at longer distances across the discourse. As the repertoire of languages explored through the lens of the NSP continued to expand between the 1980s – 2000s,

---

5 This view is by no means absolutist, Camacho (2013:112), follows Cole (2009), in the that:

> […] grammars set the minimal level of morphological specification at which null subjects can be recovered, and this level ranges from full ϕ-feature specification (person, number and gender) as in Tarifit, to no morphological specification whatsoever as in Chinese.

Their proposal for a "recoverability scale" (Camacho 2013:113) or "morphological maximality" (Cole 2009) is revisited in Chapters 4 and 7. Camacho (2013:152-153) adds further that: "[…] even in an agreement-rich language like Spanish, null subjects are not always identified by rich agreement, but by a topic antecedent."

it was noticed that many languages are better classified as **partial Null Subject Languages** (NSLs henceforth).

The partial NSL classification is usually based on limitations in the distribution of Ø subjects and non-adherence to the related morphosyntactic properties under the NSP. For example, overt pronouns may occur at unexpectedly high rates in partial NSLs and are often used in seemingly functionally redundant or infelicitous contexts (cf. Duarte 1993, 1995; Toribio 1993, 2000; Barbosa, Kato & Duarte 2005; Bullock & Toribio 2009; Sessarego & Gutiérrez-Rexach 2017; *inter alia*). In other partial NSLs, the availability of *pro* may be limited to only certain person-number configurations (Vainikka & Levy 1999; Holmberg *et al.* 2009; Camacho & Elías-Ulloa 2010; Wratil 2011; *inter alia*), to certain levels of clausal embedding (Vainikka & Levy 1999; Modesto 2008; Holmberg *et al.* 2009; Costa & Pratas 2013; *inter alia*), or is restricted by the semantic referential properties of its antecedent (Cardinaletti & Starke 1999:175-176; Modesto 2008; Holmberg *et al.* 2009; Costa & Pratas 2013; *inter alia*). Another characteristic of partial NSLs is that one does not observe all the properties usually assumed to cluster with the availability of *pro* under the NSP (Gilligan 1987; Toribio 1993, 2000; Kato & Duarte 2003, 2005; Barbosa, Kato & Duarte 2005; Odróñez & Olarrea 2006; Roberts & Holmberg 2010; Sessarego & Rodríguez-Riccelli 2018; *inter alia*).

These realizations led several researchers to reconsider the validity of the NSP and Parametric approaches more generally; since early conceptions of 'Parameters' were binary in nature, the existence of 'in-between' languages such as partial NSLs required various ways of reframing the model to account for typological variation. To name just a couple examples, Modesto (2000, 2008) proposed that Brazilian Portuguese exhibits the properties of a topic-prominent language in that embedded *pro* is identified by subject-oriented topics via A-bar topic chains (see Camacho 2008, 2013:34-66,112-115,146-173; 2016:28-30,31-38; and references therein, for detailed discussion of all of the above properties of partial-NSLs). Recently, Duarte & Soares da Silva (2016) provided data from several European and American Romance varieties that support a gradient conception of the NSP, whereby consistent NSLs like Italian lie at one extreme

of a NSP continuum, and partial NSLs like Brazilian Portuguese lie at the opposite end of the continuum with respect to their adherence to properties described above[6].

Despite the many advancements in accounts of partial NSLs within the NSP framework, there are still many aspects of Ø subjects that have not been acknowledged or have not received sufficient attention from those working in this research program. For instance, Wratil (2011), Posio (2012), and Torres Cacoullos & Travis (2019) note that overall rates of null subjects vary dramatically across languages, but that these rates ultimately reveal little about the discursive contexts, probabilistic constraints, and diachronic processes that exert an influence on the outcome of null/overt SPE. Indeed, a fully descriptive account of the selection of overt pronouns vs. Ø in partial NSLs and consistent NSLs alike eludes descriptive adequacy without some understanding of the probabilistic effects exerted by a confluence of independent predictors (Otheguy, Zentella, & Livert 2007; Martínez Sanz 2011; Otheguy & Zentella 2012; Carvalho, Orozco, & Lapidus Shin eds. 2015; Duarte & Soares da Silva 2016; Claes 2017; and references therein). Furthermore, even in languages like English that have traditionally been classified as non-NSLs, Ø subjects are nonetheless well-attested (5) (Haegeman 1990).

(5) *A very sensible day yesterday. Ø saw no one. Ø took the bus to Southwark Bridge. Ø walked along Thames Street.*

In the usage-based/functionalist/'cognitive linguistic'/typological tradition, and in the adjacent subfield of variationist sociolinguistics, the availability of Ø subjects is not tied to a underlying morpheme (*pro*), a set of correlating morphosyntactic properties (the NSP cluster), featural specifications, nor other formal operations assumed to be at the core of the language faculty under innatist approaches. Perhaps as a consequence of this theoretical outlook, anaphoric or contextually identified Ø subjects like in (5) have been explored more thoroughly under these

[6] See Quint (2008b) for an analysis of null subjects, subject clitics, and person-number inflection across Romance varieties as reflecting core vs. peripheral time-space orientation relative to each varieties' diachronic latinate source.

approaches. There has also been a greater willingness to explore Ø subjects in languages that under the NSP were considered to be non-NSLs; perhaps because the realization of anaphoric Ø is viewed as arising from the coalescence of several probabilistic constraints (rather than formal criteria), there has been less of a need to limit the typological scope of languages under examination.

The constraints that have been found to be determinative for SPE or **anaphora resolution** under the 'Probabilistic Grammar'[7] approaches are related to the **accessibility of discourse antecedents**. Accessibility is itself a property that implicates several interconnected aspects of cognition: domain-general cognitive processes, the semantic-conceptual system, the language specific production and processing domain where the vagaries of language structure enter into the equation, and factors external to one's self and one's faculties, but that nonetheless impact the distribution of linguistic forms, such as social categories and the nature of the communicative exchange (cf. Ariel 1990; Chafe 1994:71-81; Givón 2017:27-30; Claes 2017).

One example of a domain-general cognitive constraint is **structural priming**, sometimes called persistence or preservation. Priming causes linguistic forms to be repeated across a discourse in contexts where there is some contextual congruency between the site where a form is first uttered and the site where it is later resumed (see Szmrecsanyi 2005, 2006; Pickering & Ferreira 2008; Torres Cacoullos & Travis 2018; for reviews of priming across subfields of linguistics). In the case of SPE, priming has repeatedly been found to be an active constraint cross-linguistically; overt pronouns tend to trigger the use of subsequent overt pronouns and Ø subjects tend to trigger subsequent Ø subjects, particularly when in concert with other constraints such as contextual or syntactic congruence, among others (cf. Cameron 1993, Travis 2007, *inter alia* for Spanish; Wagner 2016, *inter alia* for English; Bouchard 2018, for São-Tomean Portuguese; Sankoff & Laberge 1978; Poplack 1980; Sorace *et al.* 2009; Carvalho & Child 2011; Torres

[7] Following Claes' (2017) call for a 'Probabilistic Grammar' research program that combines perspectives from variationist sociolinguistics with usage-based, functionalist, and 'cognitive linguistic' approaches to morphosyntax, I will use the unifying term 'Probabilistic Grammar' to refer to these related disciplines.

Cacoullos & Travis 2016; Cerrón-Palomino 2018; Sodaci, Backus, & Kootstra 2019, *inter alia*, for various languages in contact).

The discourse status and morphosyntactic encoding of a subject referent also determines the degree to which that referent is cognitively **accessible**; discourse referents that are most **topical**, **activated**, and/or **salient** (themselves properties determined by a confluence of overlapping factors[8]) are those that are most likely to be coreferenced by SCs over weak and strong pronominals, and by Ø anaphora over SCs and other overt pronominals (cf. Givón 1983a, 2001[1984]:399-419, 2017:3-8; Ariel 1990; Siewierska 2004:46,148-149,173-213; Speyer 2016; *inter alia*). Similarly, the way that referents are strung together across clauses in the discourse can also be determinative for SPE or anaphora resolution[9]; for example, when an anaphoric subject is in a clause that is syntactically and/or prosodically linked to the immediately prior adjacent clause containing its antecedent, Ø will be more likely to occur than overt subject forms (cf. Torres Cacoullos & Travis 2014, 2015, 2016, 2019). In a grouping of chained clauses falling under a coherent intonational and thematic contour, atonic and Ø anaphora are favored over tonic pronominals in clause-chain-medial, rather than chain-initial or chain-final positions (Givón 1976, 2001[1984]:475, 2017:5-6).

As was noted, since the availability of anaphoric Ø cross-linguistically is viewed as contingent upon a range of predictive factors under probabilistic approaches, there has been more willingness to explore Ø anaphora in languages that have historically been classified and non-NSLs or that lack 'rich' inflectional morphology for identifying Ø subjects (cf. Baker 1972; Givón

8 These include the referent's person-number encoding (whether the referent is a Speech Act Participant), the syntactic role of an antecedent in its containing clause, the semantic properties of the subject referent (animacy, definiteness, specificity), the distance between an anaphor and its antecedent (anaphoric distance), whether there are competing discourse referents among antecedents (referential ambiguity), the frequency with which a discourse referent has been mentioned (entrenchment), and even shared real-world knowledge between interlocutors. Note that some of these, like anaphoric distance and frequency-of-mention, are themselves most likely driven by domain-general cognitive processes, such as those related to short-term memory, or, as is the case for priming, may be examples of implicit learning.

9 Approaches that merge concepts of structural and semantic coherence between clauses are not fully discussed and analyzed in the present study; these include Foley & Valin (1984) and Suckow & Holler (2016), among others. These approaches comprehensively combine notions of the semantics of subordination, argument structure, and the sequential and thematic organization of discourse events.

1976, 1983b, 2001[1984], 2017, 2018[1979]; Li & Thompson 1976; Mufwene 1988; Syea 1993; Harvie 1998; Bartens 2003; Wratil 2009, 2011; Bailey 2011; Sippola 2011; Torres Cacoullos & Travis 2014, 2015, 2016, 2019; Bartens & Sippola 2014; Wagner 2016; Travis & Lindstrom 2017; *inter alia*).

Nevertheless, some scholars have found ways to analyze null subjects in such inflection-less (or inflection-'poor') languages using the formal machinery of the GG tradition by appealing to explanations such as the hierarchical and formal status of SCs as syntactic heads (DeGraff 1993; Baptista 1995, 2002), conditions of finiteness and non- or quasi-argumental status (Kouwenberg 1990, 1994; Lipski 1999; Nicolis 2008), issues of register and dialectal variation (Haegeman 1990; Haegeman & Ishane 2001), or gradience in the language-specific requirements on the degree of morphological richness in featural specification necessary for *pro* to be identified (Cole 2009; Camacho 2013:112-145). These insights are crucial for understanding SPE in CVC since the realization of anaphoric Ø in this language is likely to reflect patterns found in other non-NSL, partial-NSL, or inflection-'poor' languages.

Another curious subject form in CVC arising from the availability of SCs is the double-subject construction (3). Double-subjects in CVC can be formed with a lexical DP or a tonic subject pronoun plus a SC, which together appear to constitute the nominative argument of the verb, encoding twice (on two distinct morphemes) the subject referent in seemingly redundant fashion. This construction is found in many languages that have SCs/atonic anaphoric subject pronouns (or person agreement markers). The semantic, prosodic, and morphosyntactic properties of double-subjects consisting of a lexical DP and a SC have been explored in French (cf. Culbertson 2010; and references therein). In another study on double-subjects across varieties of Portuguese, both the lexical DP + SC type and the tonic pronoun + SC type (X2SBJs) were analyzed (Tavares Silva, Carvalho, & Ziober 2018). Beyond these studies, little is yet known about the discursive functions of X2SBJs that are constructed from the combination of a tonic pronoun and a SC. This study aims to identify the discourse function of X2SBJ constructions by exploring the set of predictive factors that condition their distribution in CVC.

Finally, a major contribution from variationist sociolinguistics has been to show that morphosyntactic structures may also be modulated by language-external constraints. These may include language-external factors that index or correlate with some social category (cf. Wolfram 1969; Labov 1972, 1978; Sankoff 1973; Rickford 1975; Romaine 1980; Winford 1984; Cheshire 1987; Tagliamonte 1998; Green 2002; *inter alia*) or those concerning the data collection materials, the interactional context in which a speech sample was collected, or the identity of the interviewer (cf. Rickford 1975; Coupland 1980; Rickford & McNair-Knox 1994; Cukor-Ávila & Bailey 2001; Schilling-Estes 2004; Bailey & Tillery 2004; Scott 2016; Childs 2017; *inter alia*). Therefore, another objective of the present study is to identify which, if any, language-external constraint might be found to correlate with patterns of SPE in CVC.

It is in this way that CVC finds itself at the intersection of only-sometimes-overlapping research programs that explore the structure of the subject domain: while the NSP has been of primary concern for research into Ibero-Romance-origin vernaculars in the Americas and Europe, debates over the status of SCs has been of greater concern in research on African languages (cf. Bresnan & Mchombo 1987; Zribi-Hertz & Diagne 2002; Creissels 2005; Dione 2013; Kari 2017; *inter alia*), but also in Gallo-Romance, Rhæto-Romance, and Gallo-Italic-Romance (cf. Rizzi 1986a; Brandi & Cordin 1989; Poletto 2000; Goria 2004; De Cat 2005; Culbertson 2010; Poletto & Tortora 2016; *inter alia*). There have been few studies that have used variationist sociolinguistic, corpus-based, or other quantitative methods to explore SPE in languages that have a disjunctive tonic-atonic opposition in their overt subject pronominal inventory, and in which these can combine in X2SBJs (but see Culbertson 2010, and references therein, for the application of similar methods to French). Therefore, another major contribution of the present study is to bring to bear these methodological perspectives on one such language, CVC, and in so doing, foster greater cohesion between these lines of inquiry in the hope that the results obtained will provide a new perspective on the synchronic interplay between tonic, atonic, and phonologically vacuous subject anaphora.

When taken as a whole, the pursuit of these objectives is intended to elevate CVC - a language that remains woefully understudied - to a position of greater relevance and attention in cross-linguistic debates over language structure and change. I hope that by applying a new methodological and theoretical perspective on SPE in this language, attention to this topic will be renewed since the pioneering works by Baptista (1995, 2002), Quint (2000a,b); Pratas (2004), Lang (2012), and Costa & Pratas (2008, 2013). Given the positioning of CVC vis-à-vis African languages, Romance languages on both sides of the Atlantic, and languages historically classified as 'creoles', a fresh examination of SPE in CVC has many contributions to offer to our understanding of how all these languages have transformed as a result of the contact between peoples from the various overlapping Atlantic spheres since the colonial era.

1.2 THEORETICAL AND METHODOLOGICAL FRAMEWORK.

In order to accomplish the objectives enumerated above, I have set out to conduct an analysis of the CVC subject domain rooted in a quantitative or variationist sociolinguistic methodology. Since this is the first study to explore the distribution of subject anaphora in CVC using this methodology, the results will provide novel contributions to the ongoing theoretical debates mentioned so far. Along the lines outlined in Claes' (2017) proposal for a "Probabilistic Grammar" research program, this study unites variationist sociolinguistic approaches with the 'cognitive linguistic' perspective in the usage-based/functionalist/typology tradition. To broaden even further the dedication to interdisciplinarity across subfields of linguistics, several aspects of the methodology applied in the current study also borrow from constructs in the GG tradition. This effort is not done merely for the sake of paying homage to disparate theoretical camps; instead, I consider such an approach the best way to arrive upon the most parsimonious possible descriptive and explanatory account of morphosyntactic structure and its determination by the human language faculty, general cognition, and social relationships.

One of the theoretical underpinnings of a variationist sociolinguistic approach is an understanding that variation in the use of linguistic forms is an inherent part of natural language: language structure is composed of **non-categorical**, **quantitative**, and **probabilistic** relationships, a property known as "normal", "structured", or "**orderly heterogeneity**" (Weinreich, Labov, & Herzog 1968; Labov 1982:7). Further, variationist sociolinguistics is deeply concerned with language change and monitors the interaction between diachronic and synchronic variability in linguistic forms, always seeking to explain the why and how of where linguistic forms come from; why they have transformed in the manner that they have and how they might undergo further transformations in the future (Tagliamonte 2006:6-7). This paradigm also seeks to ascertain the social meaning of linguistic forms or to understand how social categories and relationships modulate the expression and distribution of linguistic forms. The result is that the **speech community**, rather than the individual, is the primary object of investigation (Sankoff 1988:157). From the speech community, the researcher seeks to collect samples of the **vernacular**, that is, speech in its most unmonitored state as it is used quotidianly (Labov 1972:208; Sankoff 1980:54).

The identification of a linguistic variable for investigation involves isolating a context in which several different but functionally equivalent linguistic forms can occur, or "alternative ways of saying the same thing" (Labov 1972, 2008:2). There has been much debate as to what extent this requirement should be loosened for cases of morphosyntactic rather than phonetic/phonological variation, since there may be no exact functional and semantic equivalence between syntactic forms (form-to-meaning correspondence) that compete for realization in a given context in the same clear-cut way as, say, allophonic variation. Thus restrictions on the criteria for identifying a morphosyntactic dependent variable must be loosened such that morphosyntactic forms with a **nearly-equivalent** or **comparable** functional status may be considered, so long as their functional near-equivalence/comparability can be justified by consistent theoretical principles and empirical observations in a sample (cf. Labov 1972, 1978; Lavandera 1978; Romaine 1980; Winford 1984, 1996; Cheshire 1987; Poplack 1993; Tagliamonte 2006:9-14; Buchstaller 2009; Brook 2018; *inter alia.*) Further, the form-function correspondence is thought

to be unstable (Poplack 1993:252) and it is the objective of the variationist to model this instability, identifying the contextual factors and correlations with predictor variables that appear to trigger overlap in the from-function correspondences among possible outcomes of a morphosyntactic response variable (Tagliamonte 2006:10-12).

The "principle of accountability" requires that, after having identified a dependent variable of interest, the researcher must record every observation in a given dataset in which at least one variant (outcome) for that variable could have occurred (as defined by the form-function correspondence/near-equivalencies for morphosyntactic variables). The context in which a given set of variants could hypothetically occur is known the **variable context** or **envelope of variation**; it is the situation in which all the competing variants of a linguistic variable are free to vary (Labov 1972:72; Tagliamonte 2006:12-13). The variable context must be identified empirically from patterns in the data, the possibility of (non-)occurrence of linguistic forms must be confirmed by their actual distribution in the speech sample, and any recording of a data point (token) associated with that form must be recorded in all of its variant iterations, regardless of how (un)usual, (un)expected, or (un)predictable that usage may seem. In this way, the researcher is sure to capture the full scope of variability associated with a given linguistic variable and "tap in" to reoccurring patterns of an individuals' use of linguistic variants (Tagliamonte 2006:13). Contexts in which only one variant of a variable occurs categorically (or never occurs) must be excluded.

In CVC, as will be described in detail in Chapter 5, when the envelope of variation for SPE in CVC is restricted to [+argumental] nominative anaphoric expressions, this encompasses at least three types of subject realization: SCs (1), X2SBJs (3), and Ø (4). Thus, SPE is the relevant response variable under analysis and the realizations in (1), (3), and (4) are the variants; each is assumed to be modulated by a confluence of predictive constraints representing both language-internal and -external forces. Since defining the variable context is an empirical endeavor, further refinements will be made to the envelope of variation after a preliminary examination of the distribution of (1), (3), and (4), and how they correlate with the selected predictor variables. The various predictor variables are constructed based on previous research on CVC, as well as SPE/anaphora resolution

cross-linguistically, and draw on research from the diverse theoretical paradigms mentioned in this chapter. The speech community from which speech was collected, also described in more detail in Chapter 5, was comprised of native speakers of CVC from various locales on the islands of Santiago and Maio.

Despite the primary methodological perspective adopted in the current study – a variationist sociolinguistic approach – I have already alluded to how this study relies on theoretical and methodological contributions from research in the GG tradition to build some of the predictor variables analyzed. For instance, the semantic classification of strong, weak, and clitic forms proposed in Cardinaletti & Starke (1994, 1996, 1999) will prove essential in testing for differences in the referential properties associated with antecedents resumed by each of the anaphoric expressions in (1), (3), and (4). The concept of c-command, from Government & Binding Theory (Chomsky 1981), informs the design of a predictor variable that assesses the role of binding relationships in anaphora resolution in a local domain, as opposed to in a non-local clausal configuration.

In addition to using theoretical constructs from the GG tradition to build some of the predictor variables analyzed, I also reflect on what the findings tell us about formal models like the Typology of Structural Deficiency and the NSP, and how these might be combined with 'Probabilistic' approaches. To reiterate, this combination of aspects of theory and methodology from paradigms that are often understood to be incompatible or conflicting in their epistemological outlook is not done for the sake of compromise; I adopt the view that drawing on the most effective descriptive and explanatory tools from across subfields of morphosyntactic inquiry, regardless of their epistemological disposition, will ultimately lead to the most parsimonious account of language use, the language faculty, language in society, and how these interact to influence language structure and change.

As Tarallo & Kato (2007[1989]) noted in a citation from an anonymous German linguist in 1887, this desire to transcend the arbitrary boundaries that are built-up by theoretical paradigms over time dates almost as far back as the modern linguistic tradition itself. In the 1980's, 1990's,

and 2000's, this trans-paradigmatic perspective was kept alive by researchers exploring diachronic changes in the subject domain of Brazilian Portuguese from the mid-19th century through the start of the 21st; during this time Brazilian Portuguese transitioned away from a canonical NSL to become a partial-NSL. This line of inquiry necessarily brought issues of language variation and change together with theoretical constructs from the NSP, forcing the trans-paradigmatic perspective and effectively shifting our understanding of Ø subjects in Romance and partial NSLs (Duarte 1993, 1995, 2000, 2004; Kato 2000; Barbosa, Kato, & Duarte 2005; Kato & Duarte 2003, 2005; Duarte & Soares da Silva 2016; *inter alia*). I follow Tarallo (2015[1987]), Tarallo & Kato (2007[1989]), and Duarte & Soares da Silva (2016:4) in the assertion that:

> […] empiricism and rationalism could be complementary and that to consider the [language] internal [probabilistic] factors so as to understand the source of variation could help the analyst to infer more general principles of grammar. [...] the properties usually associated with parameters, as proposed in Chomsky (1981), contribute to formulating hypotheses and specifying the factor groups as a first step to investigate the variation of forms in competition that necessarily precedes any linguistic change; in the same way, the tendencies identified by variationist research help determine, revise, and update the properties related to the parameter issue.

Along these same lines, Martínez Sanz (2011:246), in exploring restructuring in the subject domain of Dominican Spanish, outlines a similar methodological/theoretical synthesis:

> The fact that [formal] structural properties arise as particularly relevant for disentangling grammatical restructuring in subject expression argues, in my view, in favor of applying a cohesive approach to syntactic variation to arrive at an adequate description of Caribbean subject grammars. The available theoretical investigations can inform quantitative research on the structural variables that might potentially regulate these grammars and that have been overlooked in previous variationist work. In turn, taking a corpus (to which variationist methodology is applied) as a point of departure to draw theoretical conclusions will allow syntactic investigations to build accurate theories on the possibilities of subject expression afforded by Caribbean grammar. […] It is the stand taken in the present study that to accurately define these potential sites of variability is a necessary step to build syntactic theories that can explain the limits afforded by formal grammars for cross-linguistic and cross-dialectal variation.

In the past two decades, during the transition from the Principles and Parameters approach into the Minimalist Program, there has been increasing momentum for a similarly hybrid approach on the part of researchers using quantitative findings to explore the bounds of inter- and intra-

dialectal morphosyntactic variation (Wilson & Henry 1998; Adger & Smith 2005; Cornips & Corrigan 2005; Kallen 2005; Sessarego 2011, 2012, 2013, 2014, 2017, 2019; Sessarego & Gutiérrez-Rexach 2010, 2011, 2017, 2018; Gutiérrez-Rexach & Sessarego 2011, 2014a,b; Sessarego & Rodríguez Riccelli 2018; *inter alia*). This approach is useful in that it allows for a mutual reinforcement between GG and variationist approaches, each contributes insights that the other may be lacking or incapable of contributing due to methodological restrictions. For instance, GG approaches contribute insights from speakers' introspective judgements about ungrammatical forms. This then forces one to consider how those judgements are related to underlying representations, use, and production. It also highlights that some linguistic phenomena (or domains of language) may be sensitive to variation emerging from social categories and relationships to differing degrees (cf. Cornips & Corrigan 2005).

At the same time, a variationist perspective contributes to this mutually reinforcing relationship a robust descriptive account of dialectal and socially modulated variation that helps define the limits of parametric variation and identify domains of categorical, as opposed to variable production/use (Martínez Sanz 2011:6-7). By modeling the interplay between synchronic variability and diachronic change one can also account for how variable use and perception of linguistic structures with overlapping form-function correspondences can 'feedback' into underlying mental representations, simultaneously altering the individual's 'grammar' while shifting that of the speech community as well.

Moving forward, these approaches might inspire greater collaboration and dialog between researchers across subfields of linguistics, bringing our still-nascent field closer to a more comprehensive and epistemologically mature account of Language. Dialog and collaboration should make clearer how to resolve seeming contradictions between, say, modularity-oriented cognitive models in 'biolinguistics' and embodied-cognition-oriented cognitive models like those in 'Probabilistic Grammar' traditions. By extension, such a resolution of contradictions in the underpinnings of historically-competing theories of language and cognition should allow for an overarching cognitive model that considers more closely the social meaning of language and the

results of quantitative analyses from variationist sociolinguistics, as was proposed for Claes' (2017) 'Probabilistic Grammar', but without discarding the many valuable contributions from the GG tradition.

1.3 ORGANIZATION OF THIS DISSERTATION.

Chapter 1 is the introductory chapter and provided a general overview of this work. It presented the language to be analyzed – Cabo-Verdean Creole (CVC) - and the linguistic constructions of interest: a subject clitic (SC), a null subject/anaphoric zero (Ø), and a double-subject construction (X2SBJ), all possible nominative anaphoric expressions in this language. It also provided an overview of background research on Subject Pronoun Expression (SPE) and anaphora resolution from a variety of theoretical paradigms and languages, and described the theoretical framework used to analyze nominative anaphora in CVC: a variationist sociolinguistic approach informed by theoretical constructs from Generative Grammar (GG) as well as other 'Probabilistic Grammar' approaches.

Chapter 2 is an overview of the history of the Republic of Cabo Verde and the Upper Guinea group of Portuguese-based creole languages, as well as a sociocultural sketch of modern CVC. This section starts with the founding of the colonial settlement on Santiago and focuses on the social and demographic dynamics that reveal Santiago to have been an early 'prototype' for colonial era settlements elsewhere in the Atlantic world that were similarly built on socioeconomic models dependent on the trans-Atlantic Slave Trade. It also makes note of the rapid economic and demographic boom-and-bust on Santiago and the long period of stagnation that followed, lasting for most of the remainder of the colonial period. Then, the relationship between CVC and related Portuguese-lexifier creole languages is reviewed, and the analysis in Parkvall (2000), Jacobs (2010, 2012), and Jacobs & Quint (2016) is adopted. These studies provided historical and linguistic evidence that the entire Upper Guinea Coast group of Portuguese-lexifier creoles has its origin in the Santiago variety of CVC. I supplement the evidence in support of this view with

additional sources from the historical and population-genetic literatures. Some attention is given to the last half of the colonial era, and is continued through to the conclusion of that era in 20th century when the nation overcame its independence struggles in the 1960s and 70s. This latter topic serves to transition into a sketch of the modern-day sociocultural profile of Cabo Verde and Cabo Verdeans, with a focus on the role of CVC. This chapter aims to provide the language-external sociohistorical background necessary to frame the sociolinguistic context in which the synchronic analysis in the current study was conducted.

Chapter 3 is a sketch of the linguistic history of the CVC subject domain. First, the reconstruction of the subject pronoun system carried out in Quint (2000a) and Lang (2012) is discussed in detail. A general outline of the morphosyntactic features of CVC's primary source languages is conducted; these are Late Medieval/Early Classical Portuguese, Wolof, and Mandinka. Attention is given to the subject domain and adjacent syntactic domains in these languages.

Chapter 4 is a review of literature on Subject Pronoun Expression (SPE), anaphora resolution, and the research programs that have explored these topics in CVC and related languages. It begins with the overt subject pronominal inventory of CVC, which leads to a discussion of SCs and X2SBJs in CVC and other languages. Since the analysis one adopts for SCs informs how one views the status of Ø subjects for a given language, the chapter then transitions into a discussion of the status of Ø subjects in CVC, followed by an examination of anaphoric Ø subjects in a variety of related languages and under different theoretical paradigms. The chapter is concluded by synthesizing the review of the literature conducted over its course, thus framing the synopsis in anticipation of the research questions posed in the next chapter.

Chapter 5 describes the overall methodology applied in the present study. It specifies the relevant research questions and offers some hypotheses for possible outcomes. It describes the population sampled, the geographic location and locales in which data were collected, the data collection materials, and the transcription procedures. The coding and analysis of the data are

described with details about the delimitation of the variable context and the predictor factors considered, as well as the statistical procedures that were adopted.

Chapter 6 provides the results of the descriptive and inferential statistical analyses. The construction of four analyses are elaborated: the first is a preliminary analysis that served to identify contexts of non-variability and further constrain the variable context for subsequent analyses, the second is mostly based on a multinomial fixed-effects logistic regression model built according to the empirically-constrained envelope of variation, the third and fourth analyses are binomial mixed-effects logistic regressions that split the three-way dependent variable outcome into two separate models with binomial outcomes that account for random effects. For each analysis, other procedures such as random forests, conditional inference trees, stepwise comparison of nested regression models for AIC scores, and predicted probabilities, are also used to examine the data. In Chapter 6, the results are described in a technical fashion in anticipation a broader framing and contextualization in the final chapter.

Chapter 7 consists of a discussion of the results as they relate to prior debates on the subject domain in CVC, and to debates on SPE and anaphora resolution in a variety of languages and under several different theoretical frameworks. This chapter, and the volume, ends with a discussion of the limitations of the present study, some possible future directions for further inquiry into this topic, and concluding remarks.

Chapter 2: A historical overview of Cabo Verde, the origin of the Upper Guinea Creole Portuguese group of languages, and a modern-day sociocultural sketch of Cabo-Verdean Creole.

This chapter begins at the early history of Cabo Verde starting with the settlement of Santiago in the mid-15th century, followed by the emergence of an economy that was completely externally oriented towards the trans-Atlantic Slave Trade. The consequence was a boom-and-bust cycle of economic growth and decline, but also the consolidation of a new Luso-African society with strong ties to the societies of the adjacent Upper Guinea Coast mainland. In this way, Santiago was an early 'prototype' for many subsequent colonial settlements throughout the Atlantic world that were similarly reliant on the trans-Atlantic Slave Trade.

I then explore previous research on the connection between the Santiago variety of Cabo-Verdean Creole and other closely related Portuguese-lexifier languages in the region, such as Guinea-Bissau Creole and Ziguinchor Creole. Following Parkvall (2000), Jacobs (2010, 2012), and Jacobs & Quint (2016), I adopt the view that the proto-vernacular of the Santiago variety of Cabo-Verdean Creole is the source for all the Portuguese-lexifier languages of this Upper Guinea Coast group. Additional historical and population-genetic evidence is advanced in favor of this view.

After discussion of the Upper Guinea Creole Portuguese group, I return to the language-external history of Cabo Verde, beginning with the economic decline of Santiago in the 16th century and the long period of stagnation that followed on-and-off for the better part of the colonial era. The last portion of the chapter is dedicated to the liberation struggle of the mid-20th century and the post-independence period. Particular attention for this period is reserved for issues of language policy and the role of Cabo-Verdean Creole in society.

Following Jacobs (2010), I subdivide the social and political history of Santiago into four approximative periods. These four periods are: (i) The settlement of Santiago in the 1460s, until the political, military, economic, and climatological crises of the late 16th and early 17th centuries;

(ii) The long period of decline and stagnation on Santiago that was driven by its increasing economic irrelevance, military harassment from competing European powers, corruption and exploitation from functionaries, repeated cycles of famine and drought, and general neglect on the part of the royal administration in Lisbon; (iii) The *companhia* 'company' era of the 18th century when the Portuguese royal bureaucracy became interested in Santiago once again; this period resulted in a re-intensification of trans-Atlantic Slave Trade lasting until abolition of slavery in the 19th century; and, (iv) The end of the colonial period, the independence struggle, and the post-independence Republic.

2.1 THE SETTLEMENT OF SANTIAGO AND THE FOUNDATION OF *RIBEIRA GRANDE*.

Prior to the mid-15th century, as far as archaeological history and other records have demonstrated, Cabo Verde was uninhabited. The formal discovery[10] on the part of the Portuguese of the islands of Santiago, Fogo, Maio, Boa Vista, and Sal, is credited to the Genovese navigator António di Noli and Lisboan merchant Fernão Gomes, in 1460 (Domingues 2007:41). Following discovery: "o rei instituiu sobre o território cabo-verdiano um verdadeiro senhorio, com poderes administrativos, judiciais e fiscais." [the king instituted an authentic lordship over the Cabo-Verdean territory, with administrative, judicial, and fiscal powers]. (Domingues 2007:45).

Just as had been standard practice in Madeira and the Azores, Santiago was divided into *capitanias* 'captaincies'; one located in southern Santiago around the administrative capital Ribeira Grande (today Cidade Velha or simply *Sidadi*) was granted to di Noli, another in northern Santiago at *Alcatrazes* (near modern day Pedra Abaixo) was granted to explorer and *contador* 'comptroller' of Madeira, Diogo Afonso (Carreira 1972:19-21; Newitt 2005:31; Green 2006:36-37; Baleno 2007a:71-72; Domingues 2007:47-48). Populating the island was difficult for various practical

10 See Albuquerque (2007) for discussion of geographic knowledge of Cabo Verde and possible explorations of the archipelago prior to the arrival of the Portuguese explores and their associates in the mid-15th century.

reasons, including Santiago's distance from Portugal, its harsh, dry climate unfavorable to agricultural production, and its overall lack of natural resources (Baleno 2007a:70).

Despite these difficulties, the populating of Santiago began in 1462, though official efforts at attracting European settlers were met with little success. The *capitão* 'captain' of the *donatário* 'land-grant' was charged with populating the island, and later collecting taxes and fees, stimulating economic development, and they had ultimate say in criminal and civil affairs (Domingues 2007:46). The first European settlers included Genovese, Sevillians, Castilians, and Portuguese from Alentejo, Algarve, and northern regions of Portugal, as well as Africans brought from the mainland, initially of predominantly Wolof and Mandinka origin (Carreira 1972:21, 1983:45) (a more detailed discussion of the African origins of the people of Santiago is expanded upon later in the chapter).

Early economic initiatives included the production of indigo dye extracted from the *urzela* fungus, horse breeding, cotton, and the forced deportation of enslaved peoples that had been brought from the mainland, most of whom were destined for ports in Spain and Portugal or the sugar plantations of Madeira at this nascent stage of the Triangle Trade (Mota 1954:685; Carreira 1982:7; Green 2006:37-38; Correia e Silva 2007a:85-86). Horses and cotton were the two most lucrative products early-on and were Santiago's most valued exports on the African mainland where they were often exchanged for ivory, beeswax, corn, and enslaved people (Correia e Silva 2007a:87; Tuck 2012:291).

One strategy for populating Santiago, which had also been relied upon to settle the other Atlantic Islands, was to rely on *degredados* 'degenerates/the degraded ones'. The *degredados* were convicted criminals and other social outcasts that had been expelled from Portugal. Many *degredados* were *cristãos-novos* 'new Christians', practicing Jews that had been forced to convert to Catholicism and who were fleeing the persecution of the inquisition in metropolitan Portugal, though other *degredados* were political dissidents or common criminals (Lobban 1995:16,23; Green 2006).

The *morgadio* ‘land tenure’ system was used to distribute large estates to members of the nobility and the *capela* ‘chapel’ system was used to distribute lands to clergy (Carreira 1982:5). The interests of the *moradores* ‘(land owning) inhabitants’ were represented in the legislative body of the *câmara* ‘chamber’ (Domingues 2007:51), which had already come into existence by 1497, along with a hospital and a church, all institutions associated with the administrative status of *vila* and veritable signs of societal growth (Baleno 2007a:72). Taxes, fees, and the payment of salaries for functionaries was overseen by the *almoxarife* ‘tax collector’, in cooperation with the *contador* ‘accountant’, who ensured that official legislation was obeyed. Restrictions were imposed on travel to the mainland and trade in prohibited goods or *mecadorias defesas*, the profits from which were enjoyed exclusively by the Crown. (Domingues 2007:54-57).

Settlement of the archipelago occurred nearly simultaneously with exploration of the adjacent mainland, and the economic and social development of Santiago was inextricably linked to life there. Since Santiago still failed to attract the interest of European settlers in any significant numbers, the Slave Trade was seen as a means of populating the island and coercing a labor force (Carreira 1982:5-6; Green 2006:37; Domingues 2007:46). In 1466 privileges were granted to the moradores of Santiago to trade with the mainland south of Arguim; this royal decree established a tax levied on goods procured there (Carreira 1972:22-23). As António Correia e Silva explains:

> Desde o início do povoamento das ilhas e durante o século XVI, Santiago foi o porto de chegada de produtos da Europa, porto de partida para o comércio com a Costa, porto de chegada de mercadorias africanas e porto de partida desta para a Europa e mais tarde para a América [From the beginning of the populating of the islands and during the 15th century, Santiago was the port of arrival for products from Europe, the port of departure for commerce with the Coast, the port of arrival for goods from Africa, and the port of departure of these (goods) for Europe and later for the Americas.] (2007a:103).

It was this 1466 decree that established Cabo Verde as an official trading post in the region and used the prospect of financial gain from trade with the mainland as an incentive to encourage the settlement of Santiago (Domingues 2007:46).

Despite the initial encouragement, this policy soon came into conflict with another common practice of the Crown: leasing out monopoly rights in the form of *arrendamentos* ‘leases’

in order to encourage exploration of *terra incognita* 'uncharted lands' and delegate the administration of newly settled territories (Carreira 1982:7). In 1469 an *arrendamento* was awarded to Fernão Gomes for trade with the mainland in exchange for exploring 100 leagues of coast to the south each year, beginning at *Serra Leoa* 'Sierra Leone' (Carreira 1972:25-26; Green 2006:38). The issuance of this *arrendamento* caused disputes over the mainland territories open to trade with the Cabo-Verdean *moradores*.

To clarify the matter, another decree issued in 1472 attempted to limit the *moradores* of Santiago to trade in the region between Arguim and *Serra Leoa*, though it is unclear if this policy had any practical impact. More importantly, this represented the first of many royal policies that were viewed by the Cabo-Verdean *moradores* as an overreach and impediment to 'free trade' (Carreira 1972:28-29; Domingues 2007:46). Another decree in 1472 placed various limitations on trade with the mainland, among these were included: a prohibition on collaboration and trade with merchants from outside the jurisdiction of the Portuguese Empire, a requirement that Cabo-Verdeans only sail on their own ships and that those ships be captained and equipped only by subjects of the Portuguese Crown, that goods purchased on the mainland be used only on Santiago and not be resold or re-exported, and that enslaved people taken from the mainland be used to populate and be forced to toil on the island, and not be sold elsewhere (Carreira 1972:31; Horta 2000:100-101; Correia e Silva 2007a:85-86). However, Santiago is far from Portugal, and as we will see, the ability of the Crown and the royal bureaucracy to impose its will on affairs there was quite limited.

In response to the limitations of the 1472 decree, many Cabo-Verdeans became *lançados* 'castaways'. The *lançados*, sometimes also known as *tangomãos*, were rouge merchants who sought their fortune on the mainland and would often spend extended periods of time living among local societies of the mainland *Rios da Guiné* region (approximately modern-day Guinea-Bissau and the Ziguinchor region of southern Senegal). They were notorious for their evasion of the royal bureaucracy and their trade in *mercadorias defesas*. The *lançados* were primary agents in the Slave Trade, to such an extent that historian António Carreira made the observation that if the Crown's

prohibitive policies had successfully been applied, "dificilmente se povoaria o arquipélago com a rapidez verificada" [it would have been difficult for the archipelago to have been populated with the observed rapidity]. (Carreira 1972:32; Domingues 2007:70).

The *lançados'* talent for subversion found support on Santiago itself; Carreira describes how the Crown's attempts to restrict trade, "provoked and encouraged the formation of a powerful and extensive smuggling network which eventually became institutionalized" (1982:11). These ties with the mainland would become increasingly important through the 16th century and played a crucial role in the birth of not just *Badiu* - the Santiago variety of CVC - but all varieties of CVC and likely the entire Upper Guinea Coast group of Portuguese-lexifier creole languages.

During the 1470s the Kingdom of Castile was competing with Portugal for control of the Canary Islands and increasingly challenged Portuguese claims over the other Atlantic islands. These and other tensions between the two kingdoms culminated in the War of Castilian Succession in which Portugal backed Juana de Trastámara, who was married to Afonso V of Portugal. The war resolved in 1479 with the Treaty of Alcáçovas, under which Castile was granted dominion over the Canaries while Portugal was able to retain control over the remaining Atlantic islands, including Cabo Verde, as well as claims to trading rights with newly chartered mainland regions in the Gulf of Guinea (Lobban 1995:18).

In 1482 the fortress at Elmina (present-day Ghana) was built, the next year a contact with the Kingdom of Kongo (present-day Angola) was established, and the settlement of São Tomé began in 1493. The year after that, the Treaty of Tordesillas reinforced Papal support for Portuguese claims over the Atlantic islands, and purportedly defended Portuguese claims over exclusive access to the trade with an extensive region of littoral West Africa stretching from Senegambia, down the Upper Guinea Coast, along the Gulf of Guinea, and on to the Kongo river and points south. In actual practice, other European powers would, from the 16th century onward, continually challenge these claims of exclusivity. For the time being, the rapid expansion of the Portuguese maritime empire greatly impacted Santiago and fomented its development into an

important node and port-of-call in the emerging Portuguese global trade network (Lobban 1995:19; Green 2006:40-43: Baleno 2007a:71).

2.2 BOOM AND CRISIS: THE FIRST 150 YEARS OF SANTIAGO.

The first half of the 16th century, and particularly the second quarter of that century, was a period of growth and relative prosperity for Santiago. Ribeira Grande remained small during the first quarter of the 16th century; by 1513 the population of Riberia Grande was described vaguely as consisting of "58 white male 'inhabitants' (not showing families), 56 natives of Portugal, 12 priests, 4 single white women, 16 black men and 16 black women." (Carreira 1982:6; see also Brásio 1962 '*Descobrimentos*').

The economy soon began to diversify, and new products appeared on the market, many of which were destined for provisioning ships. These included salt, animal products such as hides, oils, and salted meats, foodstuffs such as sweet potato, beans, and cassava, and later the famed *panu di terra* 'cloth of the earth' - cloths weaved in the Wolof tradition with domestic cotton and colored with locally harvested dyes – that would become important units of trade in the economy of the Upper Guinea Coast (Carreira 1983:40-44). The production of sugar on Santiago would never reach the scale of São Tomé where a major sugar monoculture plantation economy emerged, primarily due to the former's dry climate, but nonetheless in the 16th century sugarcane became an important feature of the island's agriculture, mostly for distillation into *grogu,* the local cane-based *aguardente* 'spirit' (Seibert 2014).

Despite the diversification of the local economy, Santiago's prosperity remained inextricably linked to the mainland. In the 1510s the Slave Trade intensified and grew to trans-Atlantic proportions. Along with this prosperity came escalating conflict between the trade restrictions decreed by Lisbon and the commercial activity of the *moradores* on Santiago. In 1512, an *alvará* 'charter' ordered that "os escravos que vierem de todos os nossos tratos e terras de Guiné, sejam trazidos direitamente a esta nossa cidade de Lisboa," [the slaves that may come from our

trade and lands of Guiné, shall be brought directly to our city of Lisbon,] (Brásio 1962:51; Carreira 1972:38; Meintel 1984:34; Domingues 2007:57,72). Needless to say the *moradores* of Santiago disliked the new policy:

> sentidno-se altamente prejudicados com esta medida, enviaram ao Rei em data não determinada, mas próxima da do alvará, uma exposição fundamentada, pedindo para que o texto de 24 de Outubro de 1512, não se <<entenda naqueles que à dita ilha vão com mantimentos e por seus tratos trazem dela escravos>>. [feeling themselves greatly aggrieved with this measure, they sent to the King at an unspecified date, closer to (the date of) the charter, a reasoned petition, asking that the text from the 24th of October of 1512, not be << understood to apply to those who go to said island with victuals and for (the purpose of) trade, bring slaves from there>>] (Carreira 1972:38).

It remains unknown if there was a response to the complaint from the *moradores* on the part of the Crown, leaving one to assume the regulation retained legal standing (Carreira 1972:40). Regardless of the official status of this new regulation, it is clear from the denunciations found in future *regimentos* 'official instructions' that subversion of the Crown's policies persisted (Carreira 1972:40). It is also apparent in the fact that the Slave Trade grew rapidly over the remainder of the 1510s, particularly after markets in Brazil and the Caribbean opened. Forced migrations jumped from 565 enslaved people being sent from Santiago in 1513, to 1,423 in 1515. In 1517 the first *asiento* contract for the shipment of enslaved people to the Americas was issued, authorizing the forced migration of 4,000 Africans from Cabo Verde (Green 2006:70). Similar numbers are provided by Carreira based on *avaliações* 'appraisals' of the enslaved population conducted by the Crown:

> Sabe-se apenas que entre 1513 e 1516 sairam de Santiago 2,992 escravos, dos quais 1,874 avaliados em 7,011$375 réis e 1,118 não avaliados, sendo 2,675 pela Ribeira Grande, 185 por Alcatrazes, e apenas 132 pela Praia. [It has become known that between 1513 and 1516(,) 2,992 slaves, of whom 1,874 were appraised at 7,001$375 *réis*(,) and 1,118 (that) were not appraised, (there) being 2,675 in Ribeira Grande, 185 in Alcatrazes, and just 132 in Praia.] (1972:44).

Thus, from the opening days of the trans-Atlantic Slave Trade, Santiago had already been established as a major slave-based economy and a crucial entrepôt. In 1518 two *alvarás* attempted to end the trade between Santiago and the *Rios de Guiné* all together by prohibiting the *moradores*

from trading there, demanding instead that all trade be conducted with vessels specifically supervised by the Crown, and by ordering *lançados* operating in the *Rios de Guiné* to vacate and return to Lisbon with their possessions (Carreira 1972:41-44). Of course, the Crown still lacked the ability to enforce these restrictions:

> "Foi uma tentativa como muitas outras. Nenhuma delas, nem as mais drásticas, produziu efeito. Os lançados permanceram sempre surdos aos apelos e imperturbáveis às ameaças de sanções." [It was an attempt like many others. None of them, not even the most drastic, produced (any) effect. The lançados always remained deaf to the appeals and unperturbed by the threat of sanctions.] (Carreira 1972:44).

The distribution of land and power on Santiago itself would be yet another point of contention between Lisbon and local power brokers in Santiago:

> The way in which the resources of the archipelago were distributed greatly influenced the development of society. It unquestionably worked towards the impoverishment of the great mass of the settled population in favour of the Crown, and of a small fraction of the rural bourgeoisie. These resolutely opposed the Crown contractors, whose privileges directly affected their own interests, and indirectly those of the more vulnerable classes. (Carreira 1982:7).

In the first half and middle of the 16th century the Crown established a number of bureaucratic offices in Cabo Verde that collectively aimed to reduce the scope of the *donatário* model of administration and the *morgado* system of inheritance in favor of a more centralized, Lisbon-oriented system (Carreira 1983:47-49). This process can be viewed as a transformation from a quasi-feudal system into one that was inextricably linked to institutional slavery and Santiago's role as a node in the then Portuguese dominated trans-Atlantic Slave Trade (Lobban 1995:22).

For example, *corregedores* - royal functionaries installed to oversee local governing institutions on behalf of the Crown – were installed on Fogo in 1518, and on Santiago in 1534, while official royal factors were installed for the Slave Trade in 1520, and for cotton in 1533. The *corregedores* acted alongside the *almoxarife* in the oversight of trade, tariffs, taxes, and adherence to official regulations (Carreira 1972:118, 1983:48; Domingues 2007:58).

In 1532 the Bishopric of Cabo Verde was established at Ribeira Grande, which *de jure* elevated the settlement to the category of *cidade* 'city', thus formally positioning Cabo Verde at the administrative and economic center of the Portuguese colonial jurisdiction for the entire Upper Guinea Coast (Brásio 1962:235-236; Lobban 1995:20; Green 2006:74; Baleno 2007a:72). During this period, the hereditary office of *capitão* saw a decline in its actual political power until 1564 when the position was replaced with a term-limited governorship (Lobban 1995:21). The large landholding *morgado* estates would continue to be a defining feature of social and administrative reality on Santiago (Lobban 1995:23), while the *câmara* took over as the primary arbiter of local power (Domingues 2007:52).

These administrative changes seemed to have resulted in (or failed to prevent) continued, albeit gradual growth of the enslaved population on Santiago. Increases came in the number of enslaved people transited through the island as well as those toiling there permanently. Evidence of the growth in enslaved people being transited through the island comes from various sources: calculations of slaving ships arriving and departing from Santiago, contracts, census figures obtained on Santiago itself. The numbers of enslaved people arriving in Cabo Verde jumped in the second decade of the 16th century and appear to remain steady through the third decade of that century (Table 1) (Lobban 1995:26).

Table 1. Slave ships arriving in Cabo Verde in select years between the 1510s-20s (adapted from Lobban 1995:26)

Year	Number of ships recorded	Average number of slaves per ship	Average total number of slaves shipped per year
1513	7	55	382
1514	14	97	1,354
1515	16	88	1,404
1528	14	107	1,491

Records indicate that both the Americas and Europe were destinations for enslaved people sent from Santiago: in 1526 a ship departing from Seville destined for Hispaniola called port in Santiago to load 170 enslaved people, and that same year, Diego de Torres of Lisbon and Juan Fernández de Castro of Seville were engaged in a contract for the annual forced exportation of 150 enslaved people (Mota 1977:686). Curtin (1975:13) estimated that throughout this period 80% of slaves sent to the Americas passed through Cabo Verde, while Thomas (1997:71) showed that ¾ of all enslaved peoples in Lima and Arequipa had arrived there via Cabo Verde. Green (2006:75) assumes these estimates to be conservative given the data submitted in Ventura (1999:121-123) showing that, between 1544-1550, of the 252 ships legally sanctioned to send enslaved people to the Americas, 247 passed through Cabo Verde, these figures excluding ships operating clandestinely that may not have been documented in official records.

Part of Santiago's growth as an entrepôt in the trans-Atlantic Slave Trade was also its function as a place for *ladinização*. This practice involved detaining enslaved people brought from the mainland to Santiago, usually for a few months to a year, assigning them a trade, forcing them to accept baptism, and to receive basic instruction in Portuguese (Carreira 1972:259-282, 1983:49-51). *Ladinização* was just the first stage in the brutal Middle Passage, and was motivated by religious indoctrination of non-Christians, as well as economic incentives, since enslaved people that had been 'acclimatized' or 'broken' to Western cultural norms and the periodic violence of slave labor commanded a higher price in American and European markets (Carreira 1982:6; Rego 2015:22).

The total institutionalization of slavery on Santiago also resulted in 'tiered' classes of enslaved people based on their assigned trade, destined point of settlement, and perceived 'Africaness' or degree of *ladinização*. Enslaved people destined to be traded and settled elsewhere in the Atlantic were known as *escravos de comércio* 'commerce salves', those who would toil in the plantations were *escravos de trabalho* 'work slaves', while those employed in urban centers like Ribeira Grande engaged in domestic labor. Enslaved people perceived as most African or still awaiting *ladinização* were known as *escravos boçais* '~bush slaves' or *novos* 'new', those born in

Cabo Verde were known as *escravos naturais* 'natural slaves', and those that had been *ladinizado* were known as *escravos de confissão* 'slaves of confession' or *ladinos* (Lobban 1995:24).

The Church's role in the institution of *ladinização*, their control of landholdings, and the promotion of Ribeira Grande to a bishopric and *cidade* 'city', all served to solidify the influence of the clergy on local political affairs. By this time, all the major power brokers that dominated the colonial administration were in-place: the clergy, the large landholding elites of the *morgados*, the naval merchant *armador* class, and the various bureaucrats and administrators including the *câmara*, the functionaries associated with the *feitorias* 'factories', the *almoxarife*, and the *corregedor*. It was common for officials in these posts to use their positions of power to enrich themselves, secure large landholdings, trade goods, and engage in clandestine trade (Baleno 2007a:75-6).

Demographically, the people of Santiago were also beginning to 'consolidate' into a fully realized mixed-race population. While by far the majority of enslaved people that set foot on Santiago were destined for other locations, the permanent enslaved population also grew steadily, so that by the end of 16th century enslaved Africans, free people of African descent, and mixed-race people, far outnumbered white European-descended inhabitants. For the first century or so, little direct reference was made to these former two groups. Nevertheless, their presence must have been significant by mid-century, since in 1546 free people of color gained access to the *câmara* (Green 2006:85; Domingues 2007:53). That free peoples of African descent or of mixed race had accumulated sufficient financial, cultural, and political capital to wield power in the *câmara* by mid-century was a consequence of the same demographic transition described by Carreira (1972:44-45) among the Cabo-Verdean *lançados* that frequented the Upper Guinea Coast mainland: in the early days of the trade, *lançados* were mostly of white of European descent, but this quickly began to change as the number of mixed-race or African descended people involved in the Trade grew.

Precise numbers for the enslaved population toiling permanently on Cabo Verde are missing until the last quarter of the 16th century. Lobban (1995:27) summarizes the available historical references to ethnic or racial demographics of Santiago as of 1582:

> "Jean Boulege has listed a population of 600 whites and mulattos, 400 free blacks, and 5,000 slaves in Ribeira Grande[11], and George Brooks states that the population of Praia included 1,000 slaves and 200 residents; neighboring Fogo had 2,000 slaves and 300 residents. António Carreira remarks that only a mere 12.7 percent of the entire population of São Tiago and Fogo were free".

Meanwhile, Pereira (2006), citing Correia e Silva (2007a[1995]), describes the population of the interior of Santiago in 1572 as being organized into eight *freguesias* 'parishes' with a total of 12,000 people. The majority were enslaved people, but also included the large landholding *fazendeiros* 'plantation owners', the clergy of the *capela* lands, as well as maroons living among the mountain crags. This scattered picture of the demographic development of Santiago suggests that by mid-16th century Santiago could be described as a fully realized mixed-race society. In this way, Santiago, along with São Tomé in the Gulf of Guinea, were the early 'prototypes' for the societies that emerged elsewhere in the Atlantic that were similarly beholden to the economic and demographic trends of the trans-Atlantic Slave Trade.

The economic growth of Ribeira Grande, and Portuguese dominance in regional maritime exploration and commerce, was increasingly met with competition from other European rivals such as the French, Dutch, and English beginning in the second quarter of the 16th century. By the 1530s the French had already begun capturing Portuguese ships on the high seas, and in the 1540s French pirates sacked and looted various settlements throughout Cabo Verde, including poorly defended Ribeira Grande on several occasions (Baleno 2007b:182-183). The English began engaging the Upper Guinea Coast (UGC henceforth) mainland in trade the following decade, diverting resources from Portuguese merchants who up until this point had remained mostly unchallenged (Carreira 1972:127; Lobban 1995:20; Green 2006:74; Baleno 2007a:72; Correia e

[11] It is unclear if these 5,000 enslaved peoples included those destined for other locations as well as the permanent laboring population.

Silva 2007c:156-158; Dalton 2012). The French and English introduced cheap cotton and textiles on the UGC against which *pano* producers from Santiago found it nearly impossible to compete.

To make matters worse for the *moradores* of Santiago, the Crown's response to competition from their rivals was to impose even more stringent restrictions on trade with the mainland (Baleno 2007a:73, 2007b:181-182; Correia e Silva 2007c:157). The *lançados/tangomãos* operating on the UGC mainland had grown steadily in number over the course of the century, and increasingly became the primary agents facilitating the Slave Trade. To the frustration of the Portuguese Crown, the *lançados/tangomãos* were open to conducting business with other Europeans and acted as intermediaries in commerce between non-Portuguese European traders and local African merchants (Carreira 1972:47-65).

In 1580 following the War of Portuguese Succession, the Iberian Union was consummated when Hapsburg King Felipe II brought the entire Iberian Peninsula, as well as non-contiguous Spanish and Portuguese territories, under unified control. While alliance with the Spanish kingdoms afforded the Portuguese Atlantic Islands greater naval protection and resulted in reinforced fortifications for important population centers like Ribeira Grande, it also made Portuguese colonial holdings a target for intensified attacks from Spanish rivals such as the Dutch and English. Almost immediately the nature of foreign attacks on Portuguese assets transformed from pirate-like looting and raiding into state-directed full-blown military engagement with the objective of pillaging or capturing targets (Correia e Silva 2007c:158-159; Baleno 2007b:183-189).

With the start of the Eighty Years' War, the Dutch began challenging Portuguese hegemony in various locations. The specter of sacking and pillaging was a constant threat in all Portuguese West African settlements. The Dutch raided the salt pans of Maio, Sal, and Boa Vista (Correia e Silva 2007c:159) and would eventually manage to capture and temporarily hold important forts and settlements like the *feitorias* 'factories' at Arguim, Gorée, and Mina (Lobban 1995:21). Over the course of the 1580s Ribeira Grande was sacked, looted, and temporarily occupied on five separate occasions. The most famous of these was the 1585 raid led by Francis

Drake that completely razed Ribeira Grande at a time when the city had already been suffering from famine (Baleno 2007b:189-190; Meintel 1984:36; Lobban 1995:21; Green 2006:141). The construction of the *Forte Real de São Felipe* was completed in 1593 and finally offered some protection to Ribeira Grande, successfully repelling another British assault in 1596. In 1598, a Dutch raiding party landed at Praia but was unable to take Ribeira Grande (Baleno 2007b:190).

Come the turn of the 17th century, Santiago's economic importance *vis-à-vis* the mainland began to diminish as competition from European competitors, and even from metropolitan Portuguese and Spanish merchants, began to eclipse the dominance that the *morardores-armadores* from Santiago used to enjoy (Carreira 1972:131-136). A 1588 *foral* 'decree' excused ships traveling from the UGC destined for Brazil, Seville, or Lisbon, from calling port in Santiago, allowing them to pay duties in Cacheu (Guinea-Bissau) instead. Although it was yet just another in a long line of administrative impediments on the *moradores-armadores* of Santiago, this decree was detrimental enough to commerce that it should be viewed as the beginning of a prolonged decline in the status of Santiago as major entrepôt in the trans-Atlantic Trade (Carreira 1972:140; Meintel 1984:36).

Soon thereafter, in response to complaints lodged by the Cabo-Verdean *moradores-armadores*, a series of *alvarás* 'charters' attempted once again to force ships traveling from Guiné to call port in Santiago, but the damage had already been done and clandestine ships going to the Americas continued evade calling port in Santiago (Carreira 1972:155). In the broad scope of the trans-Atlantic Slave Trade, by the second decade of the 17th century, shipping patterns began to shift away from the UGC, southward towards the Gold Coast and Gulf of Guinea, so that the number of ships departing from *os Rios da Guiné* and *Cabo Verde* declined sharply after a dramatic increase at the turn of the 17th century Lobban (1995: 27) (Table 2).

Table 2. “Number of Spanish Slave Ships Departing Cape Verde and Guinea, 1551-1640.” (Lobban 1995:27).

Years	Cape Verde and Guinea	Lower Guinea Coast
1551-1570	64	0
1571-1590	66	33
1591-1610	264	300
1611-1625	48	225
1626-1640	6	216

Santiago’s decline in economic importance was coupled with a series of catastrophic famines from the 1580s through the second decade of the 17th century, caused primarily by overgrazing and tree felling which degraded the soil (Green 2006:140-141). Tobias Green describes the coalescence of factors that contributed to the decline of Santiago at the turn of the 17th century:

> The combined effect of these three prongs of Iberian political confusion, piracy and environmental collapse created a crisis from which Cabo Verde’s metropolitan and trans-Atlantic orientated economy never entirely recovered. It is worth pausing to reflect that each of these aspects was related to Cabo Verde’s externally-oriented economy: the goats were used mainly to produce hides rather than meat, for instance, while the interest of pirates in the region came because of the wealth brought to the archipelago through international trade; the extractive cycle had, apparently reached its limits, and was followed by the collapse of this externally-oriented economy (2006:141-142).

This full-blown crisis in Santiago triggered large migrations of settlers to the mainland, in particular to the Luso-African settlements known as *praças* ‘plazas’ in Cacheu and Ziguinchor. This went hand-in-hand with a shift in political, religious, and economic power to the UGC, exemplified by the establishment of the fort at Cacheu in 1589 (Green 2006:141-142). This transfer of people and power from island to mainland will become crucial to understanding how an emergent proto-Upper Guinea Creole Portuguese vernacular was dispersed regionally.

In just 150 years after the establishment of a permanent settlement on Santiago, it transformed from a far-flung outpost at the edge of the world known to Europeans, into a central nodes in an extensive trans-Atlantic trade network. Then, Cabo Verde entered a period of decline that would return it to its early status of remote outpost.

Santiago's early reliance on enslaved people from the mainland to populate the island and its economy's dependence on their forced labor resulted in dramatic demographic imbalances between the small number of inhabitants of European descent and the majority of the population who were of African or mixed descent. In this way, the sociohistorical profile of Santiago serves as a sort of prototype that would be replicated in many colonial slave-economy-based societies that emerged later in the colonial era. It also meant that the social and demographic conditions on Santiago were precisely those associated with the kind of morphosyntactic restructuring observed in many contact languages and languages historically labelled 'creoles'. In the next section, these conditions are examined more closely, and the evidence suggesting that this newly restructured vernacular was taken to the mainland are described.

2.3 The formation of the Upper Guinea Creole Portuguese group of languages.

In the past two subsections I reviewed the early history of the settlement of Santiago and the formation of new mixed-race society there. It was noted that Santiago's economy was completely oriented towards and dependent upon the trans-Atlantic Slave Trade. In this way, the economic, demographic, and social development of Santiago was an early prototype for many other settlements established throughout the Atlantic basin from where languages historically labelled 'creoles' would come to emerge. I turn now to what those sociohistorical conditions can tell us about the linguistic formation of the Upper Guinea Creole Portuguese group of languages.

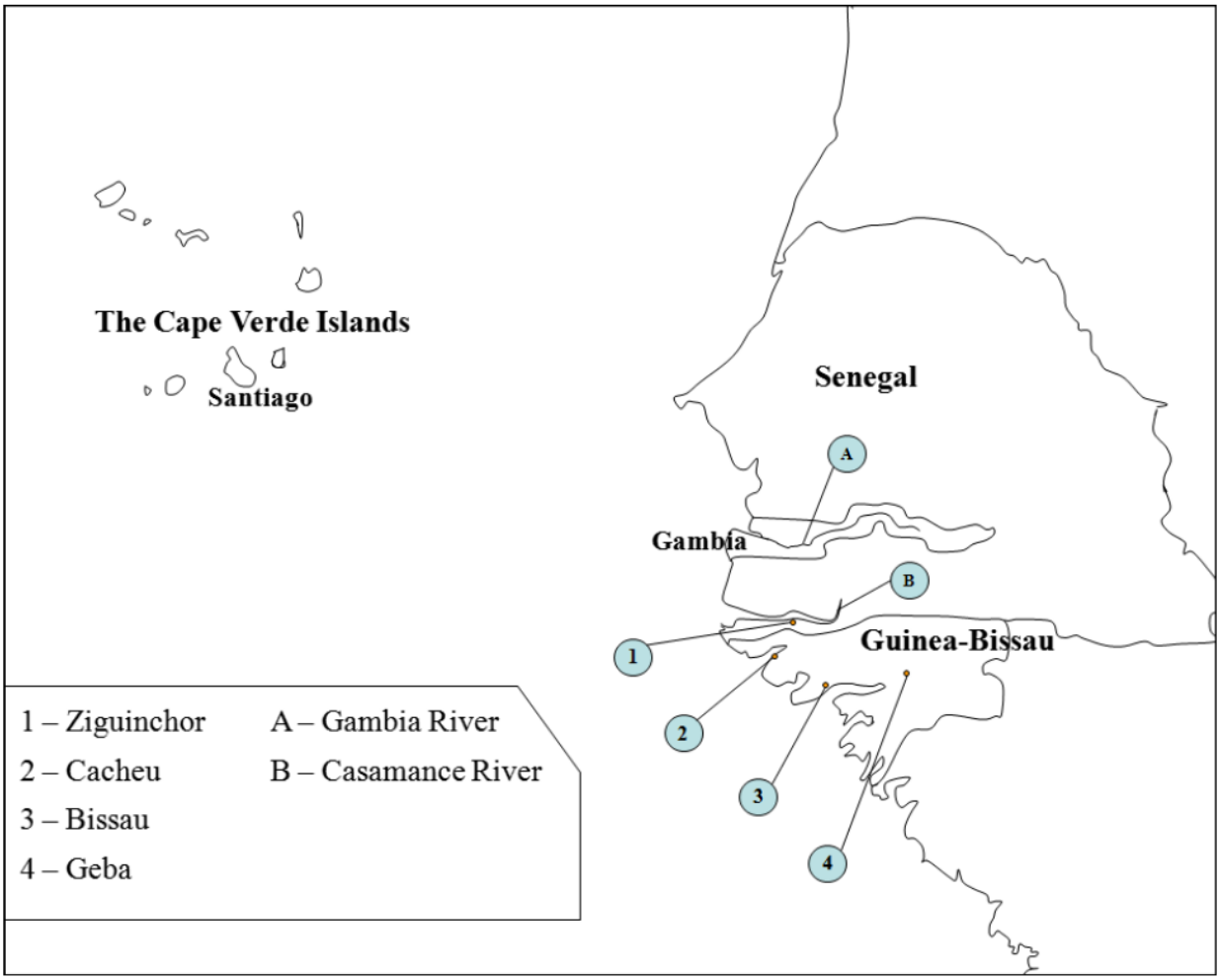
The debate over the origins of what Jacobs (2010, 2012) and Jacobs & Quint (2016) have called the Upper Guinea Creole Portuguese (UGCP henceforth) group of languages requires an examination of the 'hereditary' or 'genetic' relationship among vernaculars of the region. Some

evidence advanced in the UGCP-origins debate is extra-linguistic: it considers the sociohistorical conditions on the UGC during the early formative stages of the contact vernacular. Evidence can also come in linguistic form: morphosyntactic and phonological features that demonstrate an implicational order of change, or lexemes that appear to derive from a particular substrate or superstrate that would have been present in one, but not another region of the UGCP.

The languages of the UGCP group have traditionally been categorized as Portuguese-lexifier creoles, and include vernaculars spoken throughout the Cabo Verde archipelago, in the Republic of Guinea-Bissau, and the Ziguinchor or Cassamance region of southern Senegal[12] (Figure 1).

[12] Jacobs (2010, 2012) and Jacobs & Quint (2016), among others, advance evidence in favor of the hypothesis that Papiamentu and its regional varieties, spoken in the overseas Dutch territories of Aruba, Bonaire, and Curação, should also be included in the UGCP group.

Image 1. The Upper Guinea Coast (from Jacobs 2010:296)



The UGCP-origins debate was not initiated solely on the basis of their sharing a common lexifier language, but rather, was based on the long-acknowledged typological similarities among these languages (cf. Mota 1954). To begin with, the continental UGCP varieties approach mutually intelligibility with Badiu (Santiago CVC). The most acrolectal metropolitan/urban varieties spoken in Praia and Bissau bear the closest resemblance, while basilectal/rural varieties bear the least resemblance[13]; as has been verified by the observation of trained linguists, as well as

[13] One of my informants, a highly educated Badiu speaker from Praia, recounted a story of visiting a rural area of Fogo and experiencing significant difficulty in understanding the local vernacular of *Foguense*. Another similarly educated Badiu speaker reported the same difficulties in rural Santo Antão. Though based on informal, casual, non-

comments provided by native speakers (Rougé 1994; Quint 2000b:121). Detailed comparison of CVC and the mainland Portuguese-lexifier Creoles consistently reveal extensive similarities across several domains of language (Baptista, Mello & Suzuki 2007; Parkvall 2000:111-112; Quint 2000b:121-145), a fact that complicates the assumption that the various languages of the UGCP group could have formed independently from one another, rather than originate from a single proto-variety[14] (Quint 2000b:142-143; Jacobs 2010; Jacobs & Quint 2016).

The *simultaneous-birth hypothesis* or "*l'ambigenèse*" (Mota 1954; Do Couto 1992, 2003; Baptista 2006, 2007) states that Badiu and its mainland counterparts formed in-situ, separately from one another, but experienced parallel developmental trajectories. Quint (2000b:142-143) argues that there are too many equivalent or similar typological features (morphosyntactic, phonological, lexical correspondences, etc.) among the UGCP varieties for them to not have originated from a common proto-source vernacular (see also Baptista, Mello, & Suzuki 2007). Jacobs (2010:293) points out that the simultaneous-birth hypothesis is probably unfalsifiable since it is unclear exactly what the extent of contact between speakers of the mainland and insular varieties must have been for such dramatic typological similarities to have developed. Over the

scientific observation, extrapolating from these observations would suggest that mutual indelibility among rural varieties of CVC from different islands, even among those of the same cluster (such as Badiu and Foguense which both belong to the Sotavento group), can be limited. Thus, the typological proximity of the various rural varieties of CVC may be comparable to the typological 'distance' between urban and the rural varieties of continental UGCP.

14 Although Rougé (1994:19) argues for a less proximate 'genetic' relationship:

> […] the possibility that Kriyol would be an offshoot of Cape-Verdean has often been entertained. Although I do not share this opinion in such an extreme form, there is no doubt that Kriyol and Cape-Verdean constitute a language group vis-à-vis the other Portuguese-based creoles of Africa, i.e. São Tomense, Angolar, Principense, and Annobonese. The grammatical differences between Kriyol (GBC) and Cape-Verdean, particularly in the Tense-Aspect-Mood and determiner systems, are far-reaching enough that no assimilation appears to be feasible. Kriyol, with its varieties, and Cape-Verdean, with its varieties, do constitute a sub-group within Portuguese-based creoles – let us call it Senegambian Portuguese Creole, for instance – but they are distinct languages by all usual criteria."

Also: "Les genèses des creoles de l'archipel et du continent, en particulier dans ce que l'on appellera la «phase de pre-créolisation», sont indissociables les unes des autres. (Rougé 2006). Nevertheless, I maintain that the brunt of the evidence to be discussed ahead, and particularly the consensus in the language-external historical and anthropological literatures (cf. Rodney 1970; Brooks 2003; Hawthorne 2003; Green 2006), make clear that the mainland UGCP varieties must have their origin in an offshoot, or several offshoots, of the proto-creole vernacular that was already well underway to being fully consolidated on Santiago.

course of the next section, I follow Quint (2000b) and Jacobs (2010) in discarding the simultaneous-birth hypothesis, and instead focus UGCP-origin theories that attempt to situate the source proto-vernacular for the entire UGCP group either on Santiago, or on the UGC mainland.

2.3.1 The UGCP group must have originated from a single proto-vernacular source.

Badiu (Santiago CVC) has been described as the variety of CVC that is the oldest, the most conservative with respect to language change, the most typologically distant from Portuguese, and the one that retained the most substrate influence (Quint 2000b:108-119). Extensive correspondences between Badiu and the mainland UGCP varieties span across linguistic domains; morphosyntax, lexicon, and shared oral traditions (Rougé 1999; Parkvall 2000; Quint 2000b:121-145; Baptista, Mello, & Suzuki 2007; Jacobs 2010; Jacobs & Quint 2016). For example, with respect to morphosyntax, Quint (2000b:125-132) and Baptista, Mello, & Suzuki (2007) systematically identified several identical, or nearly identical, form-to-function correspondences in the Tense, Mood, and Aspect systems of Badiu and mainland UGCP, a portion of which are summarized in Table 3.

Table 3. TMA correspondences between Badiu and Mainland UGCP (adapted from Quint 2000b:129).

Comparison of the Tense/Aspect and Voice markers of Mainland UGCP and Badiu			
Mainland UGCP		**Badiu**	
Form	**Function**	**Form**	**Function**
ø	present / perfective (simple past) / active voice	ø	present / perfective (simple past) / active voice
-du	passive (voice)	*-du*	passive voice
ba	past (tense)	*-ba*	past (tense)
ta	habitual (aspect)	*ta*	habitual (aspect) / future (tense)
na	progressive (aspect) / future (tense)	*sata*	progressive (aspect)
iar	eventual (aspect)	*al*	potential (aspect)
ja	?[15]	*dja*	actualized (aspect)

Similar correspondences can be found in the subject pronoun inventories of Badiu and mainland UGCP (Table 4). Both systems have tonic-atonic oppositions in their pronoun inventory[16], and both encode person and number, but gender only in polite 2nd person pronouns.

15 The status of UGCP *ja* or *dja* is contested, with some evidence for its TMA particle status and contrary evidence for its adverbial status. Further, the functional value of *ja* is also contested in mainland UGCP.

16 Rougé (1994:150-163) describes the strong or tonic forms as "non-argumental" and the clitic forms as "argumental". Pratas (2002:55-61) elaborates a series of syntactic tests in favor a tripartite analysis following Cardinaletti & Starke (1994, 1996, 1999), where strong forms are disyllabic with initial *a-*, weak forms are monosyllabic lacking initial *a-*, and clitic forms are as listed in Table 4 (see Section 4.1 ahead).

Table 4. Correspondences in the pronoun inventory of Badiu and mainland UGCP (adapted from Kihm 1994:151 and Baptista, Mello, & Suzuki 2007).

	Badiu				**Guinea-Bissau Creole**			
	Subject strong	Subject clitic	Object (clitic)	Possessive	Subject strong	Subject clitic	Object (clitic)	Possessive
1SG	*[a]mi*	*n-*	*-m/-n*	*nha*	*[a]mí*	*n-*	*-n*	*nha*
2S	*[a]bo* *[a]nha* (f) (polite) *[a]nho* (m) (polite)	*[b]u-* *nha* (f) (polite) *nhu (m) (polite)*	*-u*	*bu*	*[a]bó* *anya* (f) (polite) *anyu* (m) (polite)	*bu* *nha* (f) (polite) *nhu* (m) (polite)	*-u*	*bu*
3S	*[a]el*	*e[l]-*	*-l*	*si*	*[a]el*	*el* *i-*	*-l*	*si*
1P	*[a]nos*	*nu*	*-nu*	*nos*	*[a]nós*	*no* *nu*	*-no*	*no*
2P	*[a]nhos*	*nhos*	*-nhos*	*nhos*	*[a]bós*	*bo*	*-bo*	*bo*
3P	*[a]es*	*es*	*-es*	*ses*	*[a]elis*	*e* *elis?*	*-elis*	*se*

Using Holm's & Patrick's (eds. 2007) "comparative creole chart", Baptista, Mello, & Suzuki (2007) find that Badiu and mainland UGCP share a 90% correspondence in both the surface lexical shape of morphemes and in their syntactic reflexes across the verbal and nominal domains. The extensive correspondences in morphological inventory and syntactic function across Baidu and mainland UGCP demonstrated by this line of research seem to provide compelling evidence against the simultaneous-birth hypotheses for the UGCP family.

The common pidgin ancestor hypothesis, proposed by Naro (1978), argued that Africans enslaved by the Portuguese in the early 15th century, who were taken to Europe and forced to serve as interpreters on future voyages to the western Sahara coast and Senegambia, must have developed a pidginized variety of Portuguese. This claim comes from examples of *lingua de preto* 'black speech' in dramas and other Iberian literary works dating to the 15th and 16th centuries, examples from which were reproduced in Teyssier (1959). This pidginized variety is claimed to later have been used as a *langue de reconnaissance* 'reconnaissance language' for communication between Africans and whites in Portugal, and on the UGC as early as the mid-15th century.

Presumably, then, this pidginized variety would have been used by *lançados* on the UGC mainland, and Portuguese speaking merchants living on Santiago. It could then be assumed that the pidginized variety operative on both the island and mainland would eventually have developed into separate creole languages in-situ, one spoken on Santiago, and the other on the mainland. They remained linked by their common pidgin origin and by the social interactions brought about by economic and political links between the territories. Aside from these affinities, the developmental trajectories of the creoles would presumably have diverged, leaving one to infer that only those features available in the original pidgin could possibly be maintained in the contemporary creoles.

While Naro's evidence for the existence of an early Portuguese-pidgin is credible, it seems improbable that the extensive morphosyntactic correspondences in the modern-day creoles could have emerged independently from one another, having departed from only the features available in a shared pidgin predecessor (cf. Jacobs 2010). Recall that many of the correspondences between Badiu and mainland UGCP are found in the TMA system, inflectional and derivational morphology, and other highly grammaticalized morphosyntactic reflexes. As Jacobs (2010) points out, a defining feature of pidgins is the lack of precisely these kind of functional items: TMA marking, passive morphology, conjunctions, etc. (cf. Veenstra 1996:259; Parkvall 2000:20,67; 2006:324; Bickerton 2001:104; McWhorter 2005:10; Bakker 2008:142; Muysken 2008:191; Jacobs 2010:294). Thus, I concur with Jacobs' on the common-pidgin ancestor hypothesis:

> […] judging from the complexity of the correspondences, it seems fairly safe to assume that the contact language underlying both SCV (Badiu) and GBC (Guinea-Bissau Creole) was a fully-fledged creole and that it was transmitted from one place to the other by native speakers. The reconnaissance language postulated by Naro (1978) might very well have existed, and, of course, if it did, it might, to an unknown extent, have contributed to the shaping of the proto-creole that, as argued in this paper, arose on Santiago. But this is of no concern here; paramount in the present context is rather that Naro's reconnaissance language (or any other hypothesized trade jargon or pidgin for that matter) did not give birth to GBC and SCV in two places separately (2010:294).

Another area of typological convergence between Badiu and mainland UGCP varieties can be found in its lexicon. Unsurprisingly, as much as 90% to 96% of the core lexicon of Badiu and mainland UGCP is of shared Portuguese-origin (Quint 2000b:35,136, 2008:70; Jacobs & Quint 2016), although this could be claimed to result from sharing a common superstrate language, rather than a common proto-creole origin.

The figures for the shared substrate lexicon, on the other hand, offer much more compelling evidence for a single proto-creole source for the UGCP family. Table 5 shows Parkvall's (2000:111) figures for the proportion of substrate lexical items according to the language family from which they are derived when UGCP is represented as a group, in CVC alone, in mainland UGCP varieties alone, and in their shared substrate lexicon.

Table 5. Origin of substrate lexical items in UGCP, CVC alone, mainland varieties of UGCP alone, and shared lexicon across UGCP, by language family (adapted from Parkvall 2000:111)

Source	In any UGCP variety	In CVC	In mainland UGCP	Shared
Mande	47%	65%	38%	54%
Atlantic	50%	28%	61%	42%
Other	3%	7%	1%	4%
n =	*210*	*74*	*116*	*20*

Quint (2000b:135) provides figures for the proportions of substrate-origin lexicon by specific language within the broader Atlantic and Mande families. Wolof (Atlantic) and Mandinka

(Mande) are identified as the most prominent sources for substrate lexemes in all UGCP varieties (Table 6).

Table 6. Origin of substrate lexical items Badiu, and shared between Badiu and mainland UGCP, by specific language origin (adapted from Quint 2000b:135)

	Total number of lexemes in the sample	Number of common lexemes	Percentage of common lexemes of total number of lexemes in sample
Core vocabulary	100	96	96%
Of confirmed African origin	63	41	65%
Of Mandinka origin	42	29	69%
Of Wolof origin	11	7	64%
Of Temne origin	4	3	75%
From other African languages	6	2	33%
Of probable African origin	79	21	27%

Quint (2000b:138) registers 65% of the total number of lexemes of 'confirmed African origin' as shared across UGCP, while Rougé's (1999) sample yields an even higher number of 80%[17]. I concur with Jacobs (2010) that: "Both the 90% grammatical correspondences and the shared African lexicon would constitute small miracles if the two creoles were not intimately related".

2.3.2 Sociohistorical evidence in favor of the Santiago-birth hypothesis.

If one accepts that the typological correspondences among the UGCP varieties indicate that this language group shares a common origin in a single proto-vernacular from which the other varieties of UGCP branched out, the most probable remaining possibilities are that the proto-vernacular formed on the mainland and was taken to the archipelago, or that it formed on the

[17] Parkvall's (2000:111) sample yields a much smaller proportion of shared African-origin lexemes: of his list of 210 substrate lexemes from across all UGCP varieties, he concluded that a total of 20 or 9.5% were shared between insular and mainland varieties. Thus, of the 74 CVC substrate lexemes in the list, 27% were shared with mainland UGCP, and of the 116 substrate lexemes in mainland UGCP, 17.2% were shared with insular varieties of UGCP.

archipelago and was taken to the mainland. The continental birth hypothesis argues for the former. Jacobs (2010) further divides proponents of the continental birth hypothesis into two camps. The first argues for the formation on the UGC mainland of a proto-creole among the emergent Luso-African communities whose trade networks spanned across Senegambia and the *Rios da Guiné* region (Kihm 1989:354; Rougé 1993:113; Couto 1992).

As was discussed in the language-external history section earlier in this chapter, many Cabo-Verdean merchants known as *lançados* frequented and inhabited the UGC mainland, often for extended periods of time. A closer look at the role of the *lançados* on the UGC mainland and their origins provides valuable insights for the UGCP-origins debate. According to historian George Brooks (2003:50), *lançados* hoping to do business or settle on the mainland were obliged to adhere to the tenants of the longstanding Western African tradition he calls 'landlord-stranger practices' or 'the landlord-stranger relationship':

> […] community leaders guaranteed lançados' personal safety and security of possessions, assisted them in commercial exchanges, and otherwise promoted mutual interests. Landlords refused to rent lançados more land than needed for dwellings and stores; this restriction was in accord with western African practices, whereby societies allocated rights to land use, not ownership. Denial of land or cultivation rendered lançados dependent on the landlords for food supplies, a factor of no small consequence in times of disagreement and dispute. […] By the close of fifteenth century some lançados came to be known as tangomaos, which in Portuguese and Luso-African usage connoted renegades or outcasts, as individuals who had forsaken European ways. Tangomaos manifestly found it to their advantage to integrate their lives with Africans even more closely than did lançados, for they wore African garments and protective amulets, underwent circumcision and scarification, participated in African rituals such as divination, and otherwise adopted African lifeways.

To curry favor with local community leaders[18], *lançados/tangomãos* often intermarried with local women who became known as *tangomãs* (Brooks 2003:51):

> One of the most important privileges accorded resident strangers, European as well as African, was that of consorting with local women – usually women who were related to or dependents of influential persons in the communities who sought to derive additional advantage from affiliations with strangers. Ignorant as they were of African ways, lançados

[18] The societies of the *ria coastline* region of the UGC mainland, which corresponds roughly to the tidewater regions of modern-day Guiné-Bissau, were decentralized societies largely lacking rigid social stratification and dense urban population centers (Hawthorne 2003:25-91).

> found women invaluable companions as interpreters of languages and cultures and as collaborators in commercial exchanges. African women took advantage of these circumstances for their own benefit no less than that of their male relatives, with the consequence that they came to exercise a crucial role as commercial intermediaries and culture brokers between African societies and lançados, a role that was subsequently exploited by their Luso-African children.

Additionally, local young men known as *grumetes* were indispensable business partners to the *lançados*/*tangomãos* (Brooks 2003:52-54):

> Grumetes' linguistic abilities made them invaluable in prosecuting commerce. They were recruited from many groups, notably, the Lébou, Niominka, Papel, and Biafada. They served varying lengths of time and were paid in trade goods. Some were the sons of rulers and elites, sent to work for Portuguese or Luso-Africans as apprentices in order to learn to navigate European-type vessels and gain commercial experiences. Other grumetes were domestic slaves hired out by their owners. […] Grumetes resided at or near lançados' dwellings in African communities and at their factories, forts, and other places of trade that developed over the centuries.

From the marriages of *lançados* and *tangomãs* were born the *fidju di terra* 'children of the land'. These mixed-race children were at the forefront of an emerging Luso-African merchant class (Brooks 2003:51):

> Living in African cultures and possessing some knowledge of European ways (more or less depending on the circumstances), Luso-Africans had unique potential to function as commercial and cultural intermediaries. Nonetheless, their opportunities and range of possibilities were largely determined by their mother's societies – the extent to which they inherited their mother's social ranking, rights to land use, and other prerogatives. In regard to such matters, however, there were significant differences between stratified and acephalous societies.

The new Luso-African merchant class was also fostered and influenced by the *grumetes* who worked in close, periodic, and frequent proximity to the other agents of the Luso-African merchant class (Brooks 2003:52-54):

> Just as lançados and tangomaos living among African societies adapted to their ways, grumetes shared some of the attributes of the Portuguese and Luso-Africans who employed them. They commonly wore European-style shirts and trousers. They employed Portuguese and Luso-African skills and practices, including building and repairing ships as well as constructing and furnishing dwellings, and adopted their food preferences and modes of preparation – all of which they helped diffuse widely across western Africa.

The first instantiation of the continental-birth hypothesis argues that these groups - the *lançados, tangomãos, tangomãs, fidju di terra,* and *grumetes* – were the 'agents of creolization', and the most probable linguistic community from which the source proto-vernacular or developing proto-creole must have emerged (Kihm 1989:354; Couto 1992; Rougé 1993:113).

The second instantiation of the continental birth hypothesis pinpoints the emerging Luso-African towns or trading centers known as *praças,* and *presidios* 'forts', such as those at Cacheu, Ziguinchor, Geba, and Bissau, as the primary locus of language-contact and creole formation or new vernacular formation (Rougé 2006). Founded at the end of 16th and start of the 17th century, these settlements would become the first formal – that is, officially bureaucratically recognized – *vilas* of the Portuguese Crown on the UGC mainland. The 'agents of creolization' under this instantiation of the theory would remain the same, the primary difference being the assumption that it took the social and geographic concentration of Luso-Africans into these *vilas* for the process of 'creolization' to take place.

Objections to both forms of the continental birth hypothesis are linguistic and historical in nature. From the linguistic perspective, the notion that creolization - or at least the type of language contact and change consistent with the extensive restructuring apparent in UGCP - is unlikely to have occurred on the UGC mainland between the 15th-17th centuries, if one adopts a sociohistorical, demographic, or 'ecological' approach to creolization (cf. Chaudenson 1992, 2001; Arends 1993, 2009; Mufwene 1996, 2001, 2008; Singler 2009). These approaches generally posit that linguistic creolization (or new vernacular formation), at least among the colonial European slave-based societies of the Atlantic, was contingent upon a highly specific set of language-external sociohistorical conditions. These conditions include stark demographic imbalances across various sectors of the population, such as the ratio of Europeans to Africans, the size of the mixed-race population, and the ratio of enslaved people born on the island as opposed to new arrivals, among many other factors (cf. Chaudenson 1992, 2001; Arends 1993, 2008; Singler 2009). Crucially, these imbalances result from the implementation of a slave-based

economic model, where consequently large portions of the population are subjugated under the administrative apparatus of a minority (Arends 2008).

These demographic and sociopolitical circumstances are determinant variables within the local 'language ecology' in which competing features of speakers' linguistic repertoires are subjected to the dynamic processes at work in all natural languages, including those related to community-level and individual-level variation, language acquisition, language contact, and other language-internal processes such as grammaticalization (Mufwene 1996, 2001). As a result, a process of linguistic restructuring can occur whereby the early superstrate vernacular is transformed, with additional contributions coming from the substrate languages, or whereby the features of the vernaculars used by the 'founder' population are privileged over competing features from other vernaculars in the 'feature pool'.

In the case of the so-called Atlantic Creoles, these processes were usually catalyzed by the transition of a colonial economy from a 'homestead' to a 'plantation' society. During the transition, the use of slave labor for economic growth increases dramatically, and the slave population reaches parity with, and eventually overtakes, the European or locally born population of European descent (Chaudenson 1992, 2001; Arends 2009; Singler 2009). This demographic imbalance interacts with the dynamic processes active in the local linguistic ecology (acquisition processes, inter-generational transmission, social networks, 'the founder's effect', etc.), presumably triggering massive morphosyntactic restructuring in the emergent local vernacular.

However, in the case if the UGC, the rich body of historical evidence is clear: from the 15th century to as late as the 19th century, European populations on the UGC mainland never achieved sufficient numbers, nor were they dispersed widely enough, nor did European or Eurafrican populations ever sustain sufficient suzerainty over local populations, for there to have been present the necessary sociohistorical conditions for linguistic creolization to have proceeded in-situ (cf. Duffy 1962; Boxer 1991[1969]; Rodney 1970; Curtin 1975; Mota 1977; Carreira 1982; Lang 2001; Jacobs 2010). Furthermore, it was noted that *lançados*/*tangomãos* went to great extents to assimilate into the cultural norms of local societies on the UGC mainland, not only by

intermarrying with local women and working with the *grumetes*, but also by adopting local patterns of dress, adornment, and other cultural practices such as circumcision (Mark 1999; Brooks 2003:50; Hawthorne 2003). Indeed, these were the terms of long standing 'landlord-stranger practices' with which *lançados*/*tangomãos* were obligated to comply if they were to do business with and reside among the societies of the *ria coastline,* which is the topographic area consisting of a narrow littoral strip of swampy estuaries that forms part of the greater UGC mangrove forests region (Brooks 2003).

As Jacobs (2010) correctly contends, under these conditions, Portuguese could hardly have been said to be a superstrate language, since the entire sociopolitical and cultural apparatus was dominated by preexisting African practices. This was even the case in the major entrepôt trading centers such as Cacheu, Ziguinchor, and Geba, which did not become fully established until the first half of the 17th century and beyond, when, as evidence shows (see the next subsection 2.3.3), proto-UGCP must have already largely consolidated (Carreira 1983:30; Quint 2000b:143; Brooks 2003:118-120; Jacobs 2010)[19]. Furthermore - recalling the historical profile of Santiago in its first two centuries detailed above - the demographic patterns, power relationships, and economic profile of the society on Santiago are much more consistent with the kind of 'language ecology' that is known to trigger linguistic restructuring to the extent that can be observed among the UGCP varieties. Under comparable sociohistorical conditions, there repeatedly emerged new vernaculars born of contact-driven restructuring in European colonial settlements throughout the Atlantic basin (Couto 1992). Given that Santiago was the first such settlement of the modern colonial era, it might be considered a smaller-scale prototype for the typical Atlantic plantation and entrepôt colonies that would later come to exist throughout the Atlantic (Seibert 2014).

[19] With respect to extra-linguistic processes of cultural creolization, such as in the domains of music, fashion, architecture, religion, or gastronomy, it is an on-going point of debate as to whether the sociohistorical conditions required for linguistic creolization are also necessary for extra-linguistic processes of cultural creolization to obtain, and whether or not these conditions ever obtained in the various mainland regions of Lusophone Africa. (cf. Seibert 2014 vs. Thornton 1998; Heywood & Thornton 2007).

Pereira (2006) comes to a similar conclusion, at least regarding the rural sector of Santiago in the 15th and 16th centuries. She assesses the implications that the findings in the historical volumes of the *História Geral de Cabo Verde* 'General History of Cabo Verde' have for our understanding of the sociolinguistic dynamics shaping the 'language ecology', linguistic input, and other sociolinguistic norms among enslaved communities living in Santiago during the first 150 years after settlement. In particular, she emphasizes the differences that must have existed for enslaved people living in the urban environs of Ribeira Grande as opposed to the rural interior.

In the former context, enslaved people lived in close contact with whites, likely numbering just one or two per slave-owning household. They were trained directly by the slave-owner in urban skills, crafts, and trades, since the overall goal of urban slavery in Ribeira Grande was to meet the demand stemming from the lack of an artisan workforce. On the outskirts of Ribeira Grande, some free people of African descent would have lived alongside poor whites. Generally, miscegenation was prevalent, particularly given the gender imbalance between white males and African descended females that was characteristic of the early colonial period. Thus, Perreira (2006) concludes that enslaved people living in Ribeira Grande likely received ample and extensive linguistic input from superstrate speakers, though other linguistic inputs available in the urban surroundings would have included many "interlanguages" as well as African languages. As Pereira suggests, these are the classic conditions favorable to 'language shift', rather than the creolization or new-vernacular formation we are concerned with here.

In the same chapter, however, Pereira describes the social dynamics with which enslaved people living on the rural *engenhos* 'sugar cane mills' contended. Citing Correia e Silva (2007b[1995]), Pereira explains that the interior of Santiago in 1572 was organized into eight *freguesias* 'parishes' with a population of 12,000, the majority of whom would have been enslaved people, but also included a small number of *fazendeiros* 'plantation owners', clergy of the *capela* 'chapel' lands, and 'maroons' living in the mountains. Recall the quote from Lobban (1995:27), recast here for convenience, on the demographics of Santiago in 1582:

> Jean Boulege has listed a population of 600 whites and mulattos, 400 free blacks, and 5,000 slaves in Ribeira Grande[20], and George Brooks states that the population of Praia included 1,000 slaves and 200 residents; neighboring Fogo had 2,000 slaves and 300 residents. António Carreira remarks that only a mere 12.7 percent of the entire population of São Tiago and Fogo were free.

From this vague picture we can assume that by the last quarter of the 16th century (the earliest period for which we have primary source evidence of census counts of the enslaved population permanently residing in Cabo Verde), the enslaved population was a substantial majority of the total, and most lived in the rural interior. The population structure would have been organized into dispersed clusters, centered around the small *cachoeiras* 'streams' that keep fertile the little arable land there is on Santiago. Each cluster, usually a *fazenda* 'plantation', *engenho* 'mill', or *capela* 'chapel' plot, usually consisted of a dozen, to as many 20 – 30 enslaved people. Each cluster was extremely isolated, as the jagged mountains of Santiago's interior create steep canyon-like valleys that are difficult to reach even from adjacent valleys. Interactions with the *fazendeiros* would have been minimal, and thus linguistic input from the superstrate would have been minimal too. Input would have come at infrequent intervals, particularly since so many of the slave-drivers were coerced from the ranks of the enslaved themselves. Therefore, I concur with Pereira (2006:172) that:

> […] a primeira fase de concentração de escravos no interior […] corresponde à imagem clássica da formação rápida dos crioulos […] [(…) the first phase of concentration of slaves in the interior (…) corresponds with the classic example of the rapid formation of creole languages.]

It therefore seems warranted to conclude that the sociohistorical conditions present on the UGC mainland in the 15th and 16th centuries were such that new vernacular formation (or creolization if one prefers) was highly unlikely; on Santiago, on the other hand, the sociohistorical context was precisely that which is known to have resulted in new vernacular formation, repeatedly, and in disparate locations, over the roughly three to four centuries of the colonial era.

[20] In the same caveat described in footnote number 3: it is unclear if these 5,000 enslaved peoples included those destined for other locations or just the permanent laboring population.

Regional population movements on the UGC also provide clues as to the formation and areal dispersion of proto-UGCP. Several historians observe that the mariners, merchants, and settlers that had begun arriving on the *ria coastline* region of the UGC mainland were primarily *santiaguenses*, Cabo-Verdeans from Santiago (cf. Rondey 1970; Mota 1977; Brooks 2003[1993]; Horta 2000; Mark 2002; Hawthorne 2003; Havik 2004; *inter alia*). Previously, *santiaguenses* had formed the ranks of the *lançados/tangomãos*, and after years of collaboration, trade, and cohabitation, they established more permanent *praças,* or small urban settlements, along the major river ways where there had previously been only trading posts.

The first was Cacheu, established at the end of the 16th century, then in the 17th century came Ziguinchor, Farim, Geba and Bissau. As was noted prior, these settlements are recognized as having been important centers for the dispersion and diffusion of UGCP throughout the region (Carreira 1983; Couto 1992; Kihm 1994; Quint 2000b). Jacobs (2010; and references therein) and Mark (1999; and references therein) cite numerous primary sources - including the journals of the Santiago-born travelers André Álvares de Almada and André Donelha who wrote in the early 17th century - that confirm that Cabo Verdeans were the primary agents who had forged alliances with local groups at key spots along the major riverways of the *ria coastline*. Since they adhered to the tenants of the 'landlord-stranger relationship', the *praças* were, with some notable exceptions, tolerated by locals that controlled the riverways, thus allowing Luso-African culture to become slowly entrenched.

By this period, when migration to the UGC mainland was peaking and official diplomatic recognition of the *praças* was obtained, Cabo-Verdean society was already more than a century and a half into the process of miscegenation and 'creolization' that would come to define Luso-African culture in the UGC region. In all likelihood, the Cabo-Verdean *lançados* engaging and collaborating with the indigenous *tangomãs* and *grumetes* already had a well-defined Luso-African identity (Horta 2000; Mark 2002). The emergent Luso-African merchant class in the context of the *ria coastline*, then, should be considered an extension of the already consolidated Luso-African community of the archipelago.

As we will see ahead, along with the consolidation of Luso-African cultural identity on Santiago came the consolidation of the UGCP proto-vernacular. In the context of the mainland, the Luso-African merchant class came to be closely identified with the use of proto-UGCP and constituted a kind of ethno-professional class (Hawthorne 2003). This strong association between vernacular, professional class, and ethnic group, meant that proto-UGCP occupied a singular 'sociolinguistic niche' on the UGC mainland that allowed this ethno-trade vernacular to flourish, be maintained, and be dispersed regionally (Rodríguez Riccelli, under revision).

A significant contributing factor to Luso-African migration from Santiago to the UGC mainland were the periodic cycles of drought and famine on the island (Mota 1977:686; Patterson 1988:303,306; Green 2006:241-243; Jacobs 2010). Enslaved people from Santiago were often freed during large-scale famines in desperation on the part of the *fazendeiro* 'plantation owner', as was also the case during the economic crises that lasted for the better part of the 17th century (Green 2006:241-24). Many of the newly freed sought a living on the UGC mainland where their cross-cultural fluency - an essential part of their Luso-African identity - allowed them to easily integrate into the ethnolinguistic patchwork of the region, and afforded them the possibility of working in the coastal and riverine trade (Mota 1977:696). In the terminology of Mufwene (1996), Cabo-Verdeans were in effect the 'founder population' for the Luso-African merchant class of the UGC. A passage from Brooks (2003:60) summarizes the role of Cabo Verdeans on the UGC mainland:

> Cabo Verde-based Portuguese, and, increasingly, Cabo Verde born Luso-Africans carried on a larger and larger proportion of Portuguese commerce with western Africa, which they called the 'Rios de Guiné do Cabo Verde' in appreciation of the rivers that gave access to the commerce of the vast hinterland. Children born in the archipelago acquired immunities and resistance to malaria, dysenteries, and other diseases endemic in western Africa. No less advantageous, they learned Crioulo, the language of the islands and the lingua franca that expedited commerce of Senegambia and the Upper Guinea Coast. Cabo Verdeans likewise acquired invaluable knowledge of numerous African languages, social institutions, and cultural practices from parents and relatives and also from captives brought to the archipelago for use in the islands and for resale and shipment to the Americas. Many men born in Cabo Verde visited western Africa for extended periods as traders, seamen, soldiers, government officials, and priests, and those returning to the archipelago often brought families and dependents with them. The evolution of Luso-African social and cultural attributes derived from the ongoing, two-way flow of people

> between western Africa and the Cabo Verde archipelago (Brooks 1993b, chap. 8). By the 1520s and perhaps decades earlier, Portuguese and Cabo Verdean mariners using sail-driven craft began to transform preexisting commercial networks. Vessels from São Tiago transported salt, cotton cloth, and other commodities to the Kolente River and other markets in Serra Leoa, anticipating the kola harvest that began in December. The changes in wind patterns and ocean currents during the rainy season that started in April enabled seamen to sail north to markets in Senegambia and the Guinea-Bissau region to barter cargoes for slaves, before returning to the Cabo Verde archipelago.

Hawthorne (2003:82) echoes this position:

> In the sixteenth century, trade from the Atlantic up the Cacheu came to be dominated by Cape Verdeans, Portuguese merchants leaving few descriptions of it [...] Cape Verdeans generally went first to one of the São Domingos tributaries, where they visited any number of the scattered trading villages collectively called São Domingos. Found there in great numbers, lançados facilitated exchanges between Cape Verdean captains and local African rulers. In the closing years of the century, Cacheu replaced São Domingos as the most important entrepôt on the Rio Cacheu […] The town began to attract increasing numbers of lançados in large part because area Papel recognized the advantages of allying themselves with them.

In this section, I reviewed body of historical research that suggests that: (i) the language-external sociohistorical context on Santiago in the 15th and 16th centuries was consistent with the conditions that obtained in several European settlements throughout the Atlantic basin where new-vernacular formation occurred repeatedly during the colonial era, and (ii) records indicate that Cabo-Verdean *lançados/tangomãos*, rather than Portuguese settlers, were the primary traders, merchants, and settlers, frequenting and living on the UGC mainland precisely when trade across the UGC was expanding, culminating at the end of the 16th century, and during the better part of the 17th century, when the Luso-African *praças* of the *ria coastline* region were established.

Crucially, the growth of the new Luso-African merchant class on the UGC was getting started at a time when a Luso-African society on Santiago had already been well established. As such - and as will be supported by linguistic evidence in the following section – a new vernacular, proto-UGCP, must have also already been mostly consolidated by this time. This means that the founder population for the mainland Luso-African settlements consisted mostly of speakers of proto-UGCP from Santiago, as is consistent with their already well defined Luso-African ethnic/professional identity.

Following Jacobs (2010), I do not posit any subsequent stage of restructuring for proto-UGCP in its developmental trajectory towards contemporary UGCP. Instead, I argue that the development of the mainland UGCP varieties is best viewed as a process of language maintenance, followed by gradual regional dispersion, and eventual typological divergence and diversification manifested in the dialectal variation among the various insular and mainland varieties of the UGCP group today.

This process of language maintenance, I argue, was driven by the close association of proto-UGCP with a singular 'sociolinguistic niche'. This 'sociolinguistic niche' consisted of a close interrelation between the Luso-African language (proto-UGCP from Santiago), other cultural practices of *santiaguense*-descended Luso-Africans (syncretism, cross-cultural brokers), their presence as a new Luso-African (mixed-race) ethnic group on the UGC, and their role in the local economy as a merchant professional class (facilitating trade between Europeans and various African groups); this 'sociolinguistic niche' allowed for the maintenance of proto-UGCP and its eventual regional dispersion (Rodríguez Riccelli, under revision).

2.3.3 Connecting the linguistic and extra-linguistic: lexical evidence in favor of the Santiago-birth hypothesis.

The linguistic evidence against the continental birth hypothesis is drawn primarily from the presence and distribution of both superstrate and substrate features across varieties of UGCP. Several linguists have used these features as a sort of timestamp that allows one to identify the site of formation for proto-UGCP and trace its dissemination across time and space.

In one recent example, Jacobs & Quint (2016) use the antiquity of several superstrate features to attempt to date the approximate period during which proto-UGCP must have been formed or consolidated. They consider first a number of etyma, present in insular and mainland UGCP varieties alike, that were widespread in Classical Portuguese (15th and 16th century Portuguese), but that fell out of use by the turn of the 17th century (Table 7).

Table 7. Classical Portuguese etyma in UGCP (Jacobs & Quint 2016)

Classical Protuguese	Cabo-Verdean Creole	Guinea-Bissau Creole	Modern Portuguese
alimária 'animal, mammal'	*lumária* 'id.'	*limária* 'id.'	*animal* 'animal'
dia *domingo* 'Sunday'	***diâ*** *dimingu* 'id.'	***dia*** *dumingu* 'id.'	*domingo* 'id.'
chẽo ['ʧẽo] 'full'	*txeu* ['ʧẽw ~ 'cẽw] 'a lot, many'	*ciu* ['ciw] 'a lot, many'	*che**i**o* ['ʃɐju] 'full'
frẽo ['frẽo] 'bit (n., for horse)'	*freu* ['frẽw] 'id.'	*friu* ['friw] 'id.'	*fre**i**o* ['frɐju] 'id.'
cozĩar [kozĩ'ar] 'to cook'	*kusia* ['kusjɐ] 'id.'	No reflex from Classical Portuguese	*cozi**nh**ar* [kuzi'ɲaɾ] 'id.'
lũa ['lũ(ŋ)ɐ] 'moon'	No reflex from Classical Portuguese	*luŋa* ['luŋɐ] 'id.'	*lua* ['luɐ] 'id.'
coma 'as, like (prep.)'	*ma ~ kuma* 'that, SAY-complementizer'	*kumá* 'that, SAY-complementizer'	*com**o*** 'as, like'
ser mister 'to be necesarry'	*mesti~meste* 'to need'	*mistí* 'to want'	*precisar, ser preciso* 'to need, to be necesarry'
Deus te/vos mantenha 'may god keep you in good health (greeting formula'	*mantenha* 'greeting'	*mantenha* 'greeting'	*cumprimento* 'greeting'

As was shown in the previous section, the *praças* of the *ria coastline* region were not established until a critical contingent of the *lançados*/*tangomãos* began to amass there at the end of the 16th century, and for the duration the 17th century. This is too late of a date for these lexemes to have entered to core of proto-UGCP grammar, since by this point, they had already fallen out

of use in the purported superstrate for proto-UGCP, Classical Portuguese. On the other hand, on Santiago, the intense language contact within the stratified social hierarchy and imbalanced population structure of a slave-based society was initiated a full 130~100 years before Luso-Africans began to conglomerate in any notable number on the mainland.

What is more, even as late as the early 17th century, the prototypical conditions associated with new vernacular formation, such as a superstrate-substrate sociolinguistic dynamic, were still absent from the UGC mainland. The Portuguese, European-descended, or Christian presence on the UGC mainland, had not reached a concentration such that might have allowed them to fully skirt the requirements of 'landlord-stranger relationship', nor for there to have been any sort of suzerainty over local populations, to an extent that Portuguese might have been considered a superstrate language, even by the loosest definition of the term (cf. Brooks 2003, Hawthorne 2003). Thus, not only were the sociolinguistic circumstances associated with new vernacular formation and extensive restructuring completely absent on the UGC mainland during the period between the end of the 16th and 17th centuries when proto-UGCP would have been forming according to the various instantiation of the continental-birth hypothesis, but the presence of the etyma in all modern varieties of UGCP had already fallen out of use in Portuguese by that same period.

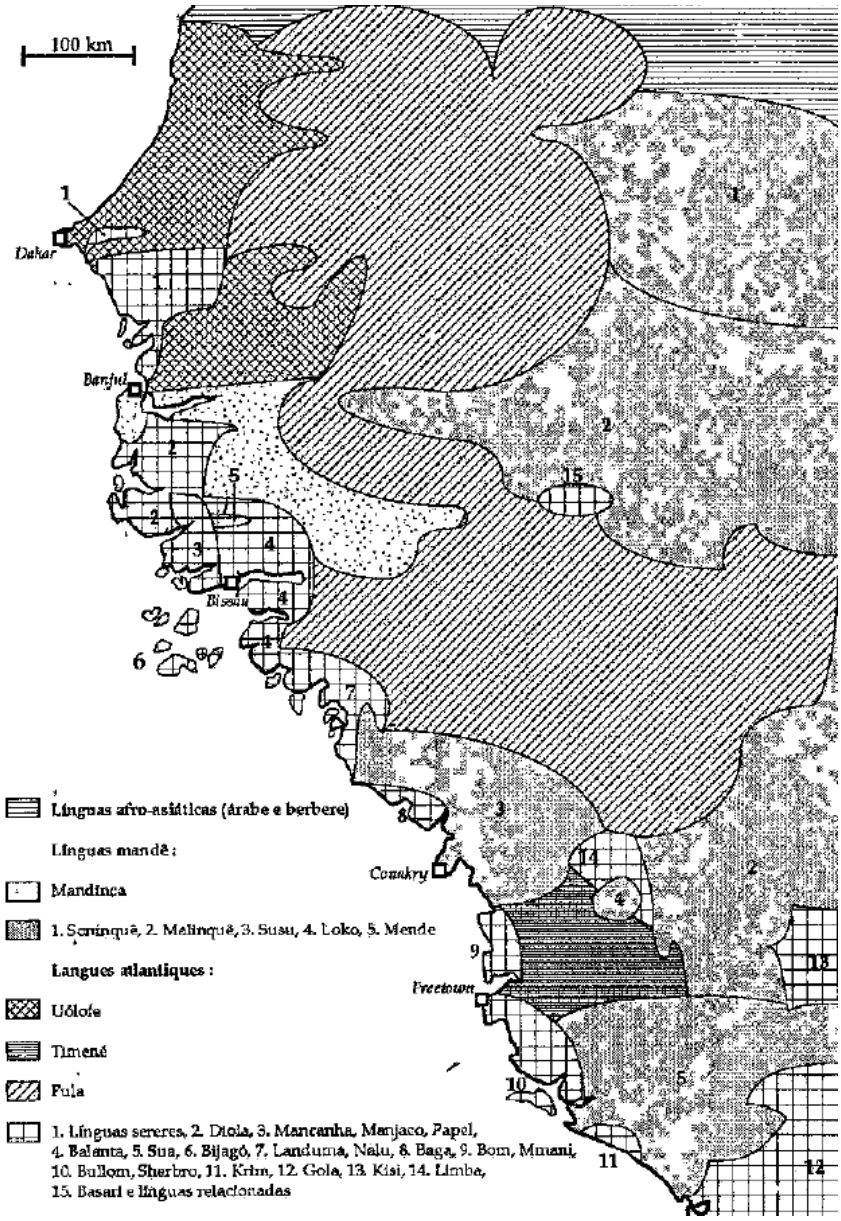
Earlier in this chapter, Table 5 showed the proportion of the shared UGCP substrate lexicon by language-family of origin (adapted from Parkvall 2000), and by specific-language origin in Table 6 (adapted from Quint 2000b). This information was used to demonstrate that the insular and mainland varieties of UGCP must have emerged from a single common ancestral proto-vernacular, since the modern-day varieties of UGCP share such a large proportion of their substrate-origin lexicon. This information once again becomes useful to us when considering more closely the areal and ethnolinguistic distribution of the contributing languages.

In Table 6, one can see that, of Parkvall's (2000:111) word list of 210 African-origin lexemes across UGCP varieties (but not necessarily shared lexemes), 47% were from a Mande language, while 50% were from an Atlantic language. Of the African-origin lexemes from across

varieties of UGCP that are shared by insular and mainland varieties alike, 54% were from a Mande language, while 42% were from an Atlantic language. As Parkvall (2000:111) points out:

> Guinea-Bissau is massively Atlantic-speaking (only about 10% of the population spoke Mande languages), so we would expect the Mande component to have been imported from elsewhere (i.e. the Cape Verdes), whereas the Atlantic adstrate influences would have boosted the already existing Atlantic share of the African-derived lexicon.

Image 2. "Languages of the Captaincy of Cabo Verde" (from Quint 2008a:24)



With respect to the absence of Mande languages in the region corresponding to modern Guinea-Bissau, and to which we have been referring as the '*ria* coastline', Quint (2008:28) argues that this approximate areal distribution would have held in the 15th and 16th centuries, at least in a geopolitical or ethno-geographic sense. This diverges somewhat from Jacobs' (2010:316) account, where, following Parkvall (2000:133), and Rougé (2006:72), he argues that Mandinka would have been a widespread vehicular language in 16th and 17th centuries. Indeed, Mandinka is known to have been widely used for trade across the greater UGC (cf. Peck 1988:85), particularly since it was the dominant language of the Kaabu Federation, the large confederation, kingdom, or empire that stretched across a large swath of interior West Africa just inland from the littoral strip of the UGC, surrounding the '*ria* coastline' to the northeast, east, and southeast. However, there are some reasons to believe that the societies of the ria coastline may have been resistant to the encroachment of the Kaabu military and the cultural and political expansionism practiced by the confederation at its borders.

A multi-ethnic state dominated by the Mandinka, Kaabu splintered off from the Malian Empire in the 14th century. Kaabu was a highly centralized, rigidly socially stratified, and militarily expansionist state (Brooks 2003[1993:46]; Barry 1998:40; Lopes 1999:76-77; Hawthorne 2003:30-32; but see Green 2009). Since their split from Mali, Kaabu had gradually annexed adjacent societies, slowly absorbing more territory at its fringes, either through military conquest or through cultural annexation via intermarriage with ruling families. This latter process of cultural assimilation was described by Mota (1954) as *mandiguização* (Hawthorne 2003:35).

Brooks (2003[1993]), Hawthorne (2003), and Havik (2004), argue that the '*ria* coastline' region was largely immune to *mandiguização*. This was possible because of two features of the '*ria* coastline' and their societies: their sociopolitical organization, the topography, and the local disease biome. The unique decentralized sociopolitical organization of the '*ria* coastline' societies (but see Nafafe 2015:82-97) allowed them to resist *mandiguização*, since Kaabu relied on exploiting preexisting social hierarchies to assimilate smaller more vulnerable societies at its fringes.

The topography of '*ria* coastline' also hindered Kaabu's military expansion; the eastern border of the '*ria* coastline' is roughly coterminous with the extent of the ocean's tidal flows, or the '*ria* coastline's' edge, a geographic border that to a certain extent demarcated the frontier between the Kaabu Federation and the '*ria* coastline' societies. Kaabu's military might came from its large horse-driven armies, but large cavalry units had difficulty operating in the dense mangrove forests and swampy web of rivers, inlets, and tributaries. What is more, the endemic *Tse-tse* fly halted cavalry advances via the spread of sleeping sickness. These factors greatly reduced the strategic threat posed by the powerful Kaabu military in the region (Brooks 2003:11; Hawthorne 2003:32-41).

Given the resistance to *mandiguização* among the societies of the '*ria* coastline', and the topographic and ecological challenges of the micro-climate, it is likely that the cultural weight of Mande as a vehicular language or regional *lingua franca* may have been less ubiquitous than previously assumed, at least within the small '*ria* coastline' region of the UGC. Recalling that this was the primary zone of settlement and trade activity for *lançados/tangomãos,* and thus the locus of Luso-African culture on the mainland, it is unlikely that Mande could have left such a large imprint of the core lexicon of UGCP having been relevant only at the fringes of the '*ria* coastline' region. The continental birth hypothesis, therefore, fails to explain how the Mande could have come to constitute such a large portion of UGCP's core lexicon. If one considers, however, the historical connections between Santiago and other parts of littoral West Africa adjacent to the '*ria* coastline', such as with Senegambia, then the Mande lexical contribution becomes apparent.

As was described in Section 2.1, in the 15th century, the mainland territory available to trade for the *moradores/armadores* of Santiago as per royal decree was much larger and included most of the Senegambia region, stretching as far south as modern-day Serra Leoa. Over the course of the next century, this region became reduced piecemeal as trade monopolies were granted by the king to individual explorers and traders. At the same time, other European powers encroached on Portuguese control of the regional maritime trade, until eventually in the mid-16th century the trade zone available to the *moradores/armadores* of Santiago became reduced to approximately

the modern-day borders of Guinea-Bissau (Carreira 1983; Quint 2000b:8). Before this shift, however, the primary regions from which enslaved peoples were taken to Santiago included regions where Mande - and CVC's other major substrate language, Wolof – exerted much greater influence, either as autochthonous languages or as vehicular languages (Silva Andrade 1996; Baptista 2002:14-16; Lang 2006; Jacobs 2010). If Santiago is taken to be the site of new vernacular formation, and the place at which a 'proto-UGCP' may have been consolidated or formed, then it becomes apparent why there should be such a heavy Mande imprint on the core lexicon of UGCP.

The origin of the core UGCP lexicon coming from Atlantic languages can be traced to the same areal shift in trade routes between Santiago and UGC mainland. Parkvall (2000:111) explains:

> […] the Atlantic component found in Guinea-Bissau PC, but not on the Cape Verdes, consists mainly of words derived from languages such as Diola, Banyun and Manjaku, all spoken in Guinea-Bissau. The shared Atlantic component, on the other hand, derives more from geographically more distant languages, such as Wolof.

Wolof is spoken (and likely was in the 15th-17th centuries was as well, see cf. Quint 2008a:28) in the portion of Senegambia north of the Gambia river, far removed from the northern most reaches of the area of the UGC targeted by the *lançados/tangomãos* at the turn of the 17th century, which ends at the Cassamance River (i.e. the '*ria* coastline'). Therefore, the Wolof portion of the core lexicon could not have entered UGCP as a substrate on the continent, but instead must have entered the UGCP lexicon as a substrate on Santiago, and at the earliest phases of the trade between Santiago and the mainland, when the region available to the *moradores/armadores* of Santiago still included the *Petite Côte* and most of Senegambia (Lang 2004, 2006, 2009, 2011; Jacobs 2010).

A similarly close analysis of the shared substrate lexicon of the UGCP varieties by Quint (2000b) leads one to draw similar conclusions. Rather than considering the language family of origin for the substrate lexeme, he identified each substrate lexeme by its specific language of origin. As shown in Table 6 above, of the 100 core-vocabulary items considered in Quint's (2000b:135) analysis, he identified 63 words in Badiu that are of confirmed African-origin. Of

these, 65% (n=41) are shared across UGCP. And of these 41 shared substrate lexemes, 70% (n=29) are from Mandinka, and these constitute 69% of all Mandinka-origin lexemes found in the confirmed African-origin subsample (n=42). Of the 41 African-origin lexemes shared across UGCP, 17% (n=7) are of Wolof origin, and these constitute 64% of all Wolof-origin lexemes found in the confirmed African-origin subsample (n=11). As both Quint and Parkvall note, the dominance of Mandinka and Wolof in the shared substrate lexicon of the UGCP modern group is disproportionate to the presence of those languages in the '*ria* coastline' region in the 15th and first half of the 16th century when proto-UGCP must have been forming or becoming consolidated. Both these languages, however, are implicated as the primary substrate on Santiago during this period.

Lang's (2004, 2006, 2009, 2011) efforts in connecting archival records with linguistic evidence confirms the early Santiago-Senegambia connection as the most probable explanation for how so much of the UGCP shared lexicon could be of Mandinka and Wolof origin. As mentioned in Sections 2.1 and 2.2, Santiago was a nexus in an early predecessor to the 'Triangle Trade'; in the 15th century a trade route connected the city of Valencia on the Iberian Peninsula to the portion of Senegambia known as the Petite Côte (in modern-day Senegal, on the mainland adjacent the island of Gorée and running south to approximately the delta of the Saloum river), via Santiago. Using records compiled by Vicenta Cortés Alonso (1964), who documented the ethnic origin of enslaved peoples arriving at the port of Valencia between 1479-1519, Lang notes that Wolofs accounted for 2,258 individuals, or 92% of the enslaved people registered as having entered through the port. Several quotes in Cortés Alonso (1964) make clear that traders and merchants operating in the port of Valencia coordinated the trade through the *feitoria* on Santiago (Lang 2006, 2011).

Furthermore, in the same volumes, Lang observes that the Cabo-Verdean travelers who had documented their experiences and observations during their time on the UGC mainland all bemoaned the encroachment of the French and Dutch on the exclusive regional privileges Santiago once enjoyed in trade with Wolof kingdoms and the loss in purchasing power of goods from

Santiago such as horses. These accounts confirm that Wolof was the predominant ethnicity of the enslaved people toiling on or being sent through Santiago in at least the first 50-100 years following settlement and the opening of the Slave Trade with the mainland. Only much later in the first half of the 16th century, when the mainland region open to trade with Santiago had already shifted southward, would the primary settlements of the '*ria* coastline' start to appear, leaving Wolof speakers far outside geographic space where the continental birth hypothesis would place the site of formation for a new proto-UGCP vernacular.

Thus, the heavy presence of both Mandinka and Wolof origin lexemes shared across UGCP varieties points to a Santiago-origin for the formation a new proto-UGCP vernacular variety from which the other varieties of UGCP branched-off and dispersed throughout the Cabo Verde archipelago, and throughout the UGC mainland (Lang 2006; Jacobs 2010). The presence of the Classical Portuguese features discussed above, the distribution of Mande and Wolof lexemes in the shared core lexicon of the UGCP group, the chronology of population movements, trade relations between Santiago and UGC mainland, and the emergence of Luso-African settlements on the '*ria* coastline', all indicate that the new vernacular emerging on Santiago - perhaps best labelled 'proto-UGCP' following Jacobs (2010) - must have been largely consolidated by the mid-16th century at the latest, or about a century after initial settlement (Jacobs & Quint 2016). In the next section, I review evidence from the domains of morphosyntax and phonology that lend further support to this notion.

2.3.4. Morphosyntactic evidence in favor of the Santiago-birth hypothesis.

Just as the presence of lexemes from Wolof in UGCP hints at the antiquity of Wolof as a substrate for proto-UGCP, so too do the traces of Wolof morphosyntactic reflexes in modern varieties of UGCP (cf. Jacobs 2010). In situations of language contact, residual morphosyntactic reflexes from one of the substrate languages still present in the contemporary vernacular usually indicate an older, 'core' substratal layer of influence (cf. Hagemeijer 2011), and that the

contributing substrate language must have been present in the 'pool of variation' at the early formative stages of the new contact-vernacular. A comprehensive review of all the purported morphosyntactic reflexes that have been attributed to Wolof or Mandinka in UGCP is well beyond the scope of the current study (see Rougé 1994, 1999, 2006; Quint 2000a,b, 2006, 2008; Lang 2003, 2005, 2006, 2009, 2011; Baptista 2006; Bartens 2006; Couto & Souza 2006; Baptista, Mello, & Suzuki 2007, for lengthy expositions of postulated Wolof and other substrate morphosyntactic traces in Badiu). For the sake of exposition, a few of the purported Wolof morphosyntactic reflexes in UGCP are discussed here.

In one example, Jacobs (2010) points to the observations in Quint (2000b:66) and Lang (2006:53, 2011:69), that the variation in UGCP between possessive verb *ten*, and the vowel-final form *teni*, both derived from Portuguese *tem* 'to have', has its origins in a convergence between a superstrate lexeme and substrate morphosyntactic features. In Badiu (and CVC generally), *ten* denotes 'essential possession', while the latter vowel-final form *teni* indicates 'occasional possession', as in to have something available, at one's disposition, or temporarily in one's physical possession (Quint 2000b:66). This is mirrored by variation in Wolof between possessive *am* and *ame*, which respectively bear the same subtle semantic distinctions as *ten* and *teni*. Additionally, Quint (2000b:66) points out that both *ten* (6) and *am* (7) can denote 'the passage of time'.

(6) *Bu ta tem dés diâ la bu ta kunsa bem*

'You were there for ten days before you started to come.'

(7) *Am nañu fi ñaari ayubés*

'It's been two weeks since they've been here'/'They've been here for two weeks'

Another purported Wolof morphosyntactic reflex is the UGCP passive construction, which lacks an auxiliary verb, and simply attaches morpheme *-du* to a transitive verb (8).

(8) *Ali tanbe Nu nxinadu Kriolu*

Here also 1.PL teach.PASS Kriolu

'Here we are also taught Kriolu'

Like *ten(i)/am(e)*, it appears *-du* may be traced to convergence between a superstrate lexeme and a substrate morphosyntactic pattern. Like passives in many other Indo-European languages, *-du* assigns to the nominal argument in subject position the role of object or experiencer, and lexically resembles Portuguese passive morpheme *-do* /-du/. Unlike Portuguese passives, CVC *-du* employs no auxiliary verb. Baptista, Mello, & Suzuki (2007:67,68) and Baptista (2007b) identified the Wolof copula-less passive suffix *–(t)u* as the potential substrate form from which the UGCP passive is derived. Jacobs (2010) takes this to be further evidence for a Wolof imprint that entered UGCP early, via Santiago. Additionally this passivizing morpheme also appears across varieties of UGCP (Lopes da Silva 1957; Kihm 1994:243; Quint 2000a:234,235, 2008; Veiga 2000:336; Jacobs 2010, and references therein).

These were just two of the many examples of Wolof morphosyntactic traces in Badiu and UGCP generally. These robust correspondences further implicate Wolof as a major substrate for proto-UGCP, thus lending support to the Santiago-birth hypothesis. With respect to Mandinka, when I review Lang's (2012) reconstruction of the proto-CVC subject pronoun inventory, we will see that he has claimed that the CVC first-person singular subject clitic *N* is a direct borrow/imposition of the Mandinka first person singular person marker *ń*. However, it is more likely that this correspondence may be indicative of cross-linguistic convergence between superstrate and substrate forms, rather than direct substrate imposition. For the time being, I turn to the next subsection in which I review phonological evidence that provides additional support for the Santiago-birth hypothesis.

2.3.5 Phonological evidence in favor of the Santiago-birth hypothesis.

Badiu (Santiago CVC) retains the Classical Portuguese voiceless postalveolar fricative /ʃ/, while voiced postalveolar fricative /ʒ/ devoiced, thus merging with the voiceless postalveolar

fricative[21] (Quint 2000a:110). Mainland UGCP, on the other hand, depalatalized in all of these contexts, developing instead alveolar realizations (Table 8) (Jacobs 2010; see also Kihm 1994:4-5; Rougé 1994:130).

Table 8. Devoicing of Classical Portuguese /ʒ/ and maintenance of /ʃ/ in Badiu and subsequent depalatalization of /ʃ/ in Guinea-Bissau Creole (adapted from Jacobs 2010:330)

Portuguese	Badiu	Guinea-Bissau Creole
peixe	/ˈpe.ʃi/	/pis/
baixo	/ˈba.ʃu/	/bas/
deixa	/ˈde.ʃa/	/di.ˈsa/
lagartixa	/la.gaɾ.ˈti.ʃa/	/la.gaɾ.ˈti.sa/
igreja	/ˈgɾe.ʃa/	/ˈgɾi.si.ja/
viagem	/ˈbja.ʃi/	/bi.ˈjas/
hoje	/ˈo.ʃi/	/aos/
sujo	/ˈʃu.ʃu/	/su.ˈsu/

A mainland birth for UGCP would suggest that postalveolar fricatives from Portuguese first depalatalized in the mainland variety of proto-UGCP, then re-palatalized on Santiago. A much more likely developmental path is that Classical Portuguese post-alveolar fricatives retained their nearly palatal realization in proto-UGCP on Santiago, voiced post-alveolar fricatives then devoiced, and finally the palatal feature was lost in mainland UGCP (Jacobs 2010).

Badiu also maintains the contrast between voiceless postalveolar affricate /ʧ/ and voiceless postalveolar fricative /ʃ/. This is another retention from Classical Portuguese, the affricate having lost its affrication and merged with the voiceless postalveolar fricative in the 17th century[22] (Tyssier

[21] Voicing is contrastive in many of these contexts in modern Badiu. Quint (2000a) and Jacobs (2010) assert that the change in Table 8 was a regular diachronic one, and only through modern superstrate influence did the voiced allophone reenter in this phonotactic context. Jacobs (2010:328$_{n50}$) explains: "The devoicing of voiced fricatives is completely regular in UGPC. Voiced allophones are attested in Cape Verde as well as in Guinea-Bissau, but are likely to result from post-formative contact with Portuguese." Indeed, rural or basilectal varieties of Badiu tend to lack the voiced allophone (Quint 2000a,b), suggesting that this variable may be a reflex of the speaker's access to the superstrate language, and the various social factors associated with diglossic societies.

[22] The distinction is also maintained in Galician and some northern varieties of Portuguese (Jacobs & Quint 2016). Interestingly, Newitt (2015) maintains that the origin of most settlers of the Portuguese Atlantic Islands were from Northern Portuguese regions like Tras-os-Montes. However, others such as Carreira (1972, 1983) emphasize other

1980:66-68; Kihm 1994:4-5; Rougé 1994:130). Mainland UGCP retains these as voiceless palatal plosive /c/, and as we saw above, voiceless alveolar fricative /s/ (Table 9) (Jacobs & Quint 2016).

Table 9. Maintenance of postalveolar fricative and postalveolar affricate in Badiu, depalatalization of the former and retention of the latter as voiceless palatal plosive in mainland UGCP (adapted from Jacobs & Quint 2016)

Classical Port.	**CVC**	**GBC**	**Gloss**
/ʃ/			
*Pei**x**e* ['pejʃi]	*pe**x**i* ['peʃi]	*pi**s*** ['pis]	‘fish’
*cai**x**ão* [kaj'ʃɐ̃w]	*ka**x**on* [ka'ʃõ]	*Kas**o**ŋ* [ka'soŋ]	‘coffin’
/ʧ/			
***ch**umbo* ['ʧũbu]	*txunbu* ['ʧũbu ~ 'cũbu]	*cumbu* ['cũbu]	‘lead’
*fe**ch**ado* [fe'ʧadu]	*fitxádu* [fi'ʧadu]	*ficadu* [fi'cadu]	‘closed’

Similarly, Badiu maintained the voiced postalveolar affricate /ʤ/ from Classical Portuguese, which eventually deaffricated and merged with voiced postalveolar fricative /ʒ/ in contemporary Portuguese (Teyssier 1980:35; Jacobs & Quint 2016). The voiced postalveolar affricate was retained only in word-initial position, while word-internally it deaffricated and devoiced yielding /ʃ/ as in Table 9, and for mainland UGCP the word-initial value was maintained as voiced palatal plosive /ɟ/ (Table 10; Jacobs & Quint 2016).

Table 10. Maintenance of voiced postalveolar affricate in word-initial position in Badiu, retention as voiced palatal plosive UGCP (adapted from Jacobs & Quint 2016)

Classical Port	CVC	GBC	Gloss
***j**ogo* ['ʤogu]	***dj**ogu* ['ʤogu ~ 'ɟogu]	***ju**gu* ['ɟugu]	‘game’
***j**unto* ['ʤũtu]	***dj**untu* ['ʤũtu ~ 'ɟũtu]	***j**untu* ['ɟũtu]	‘together’
***g**emer* [ʤe'meɾ]	***dj**emi* ['ʤemi ~ 'ɟemi]	***j**imí* [ɟi'mi]	‘groan’

sending regions of Portugal such as Algarve and Lisboa, while Lang (2001) laments that it may be impossible to know the sending regions from where the Portuguese settlers of Santiago hailed.

Tyessier (1980) observed the voiced postalveolar affricate for the period of Old Portuguese from 1200-1350 but makes no mention of its existence in the period from 1400 through the modern day. Since a set of competing linguistics variants can linger for a significant amount of time even after one of the variants has declined in use, Jacobs & Quint (2016) assume that /ʤ/ remained an active competing variant in the feature pool of the Old Portuguese superstrate on Santiago in the 15th century. For /ʤ/ to have been present in the feature pool for proto-UGCP under a continental birth hypothesis, it would have had to have persisted in the superstrate for an additional 250 years after its last observation in Old Portuguese.

2.3.6 Evidence from population genetics in support of the Santiago-birth hypothesis.

An innovative study combined population-genetic data with data on idiolectal lexical variation (Verdu *et al*. 2017) and seems to provide further support for the Santiago-birth hypothesis. The discussion of the 'origins debate' for UGCP thus far has linked the substantial number of substrate features from Mandinka and Wolof in the lexicon and morphosyntax, to trade patterns and population movements between Santiago and mainland during the first 100-200 years after the settlement of the former. In particular, the Senegambia region, where Mandinka and Wolof were (and are) widely used, eventually became closed off to trade with Santiago. In the *ria coastline* region of the UGC - where a new Luso-African merchant class that was closely associated with the use of proto-UGCP emerged, coalescing at the end of the 16th and for the duration of the 17th centuries – Wolof, and to a lesser extent Mandinka, were not sufficiently dominant, not as indigenous languages nor vehicular languages (see Section 2.3.2), and certainly not in any proportion consistent with the substrate contribution of these languages to the shared African-origin lexicon across varieties of UGCP.

In Santiago, on the other hand, there was extensive trade with the Senegambia region early-on, and based on records from slaving vessels and accounts from travelers writing at the time, the African-descended component of the enslaved population being taken to or transited through

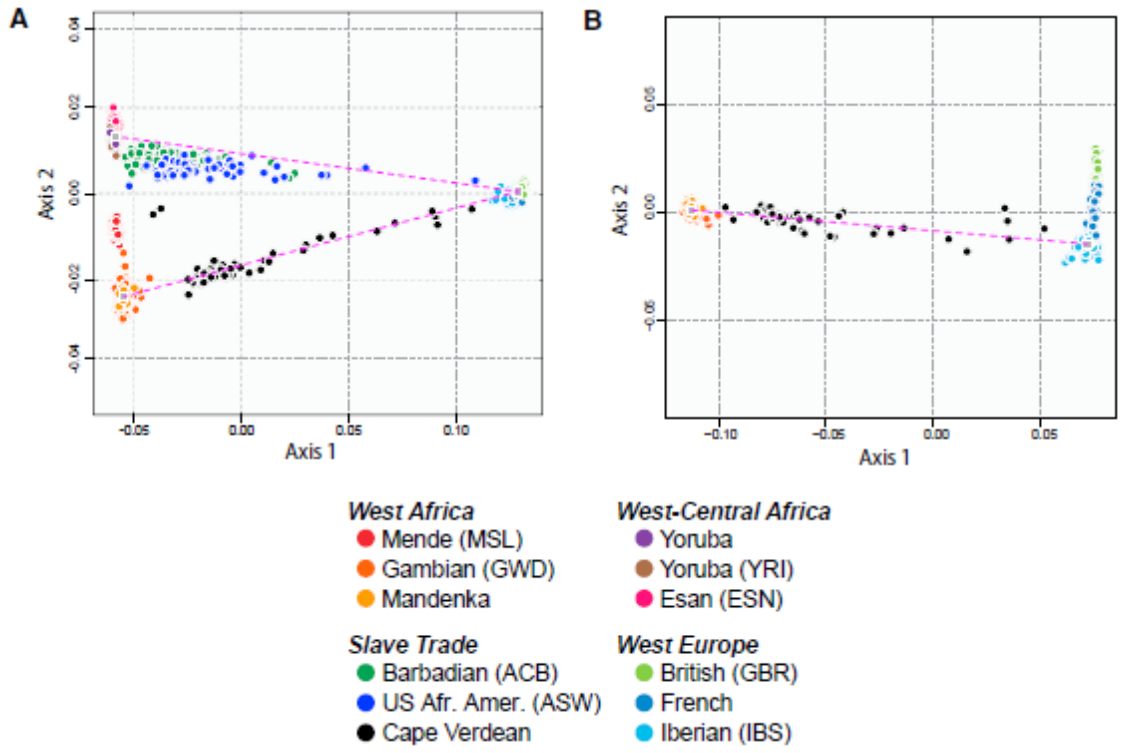
Santiago appeared to be of predominantly Senegambian origin at the opening stages of the trans-Atlantic Slave Trade and the early days of settlement of Santiago (Lang 2009, 2011). Evidence from population genetic studies further support this notion, as a significant portion of the Cabo-Verdean gene pool reflects Gambian Mandenka and Senegalese Mandinka ancestry.

Verdu *et al.* (2017) explored the relationship between genetic and idiolectal variation in a sample of the Cabo-Verdean population from across the archipelago, with the genotype samples compared with similar samples from African mainland, Europe, and other African-diasporic communities in the Atlantic zone. While the main goal in Verdu *et al.* (2017) was to link observable trends in genetic and linguistic variation, this topic is beyond the scope of this chapter. Nevertheless, the comparison of genotype samples from Cabo-Verdeans with other West African populations provides partial confirmation of the Santiago-birth hypothesis.

Figure 3 shows a "two-dimensional multidimensional scaling (MDS) analysis of individual-pairwise-allele-sharing dissimilarities", where each point represents an individual. The two MDS analyses are described in the paper as follows (Verdu *et al* 2017:2530):

> A) MDS analysis of six continental African populations from West and West-Central Africa, three Western European populations, and three admixed populations that arose during the transatlantic slave trade. The dotted lines connect the centroid of the Europeans with the West-Central African centroid and the centroid of Gambian Mandinka and Senegalese Mandenka. Each centroid is marked with a gray square. (B) MDS analysis of two West African populations (Gambian Mandinka and Senegalese Mandenka), three European populations, and the Cape Verdean population. The dotted line connects the West African centroid with the Iberian centroid.

Figure 1. “Genetic Variation in African–European Admixed Populations in Relation to Continental African and European Populations” (from Verdu *et al.* 2017:2530).



The results in Figure 3 show that Cabo Verdeans are a population resulting from admixture between Western Europeans and West Africans from the Senegambia region, namely Senegalese Mandenka and Gambian Mandinka. To reiterate, these populations are from Senegambia, and not the ‘*ria* coastline’ portion of the UGC, which is much further to the south. That so much of the Cabo-Verdean genotype can be traced to these population makes clear that speakers of Mandinka formed a substantial part of the enslaved population on Santiago. Given the chronology of trade relations between the archipelago and the mainland, we know that the majority of this population was likely brought to Santiago during the period in which trade with the mainland was open to the *moradores/armadores* of Santiago, that is, precisely the same period during which proto-UGCP was taking shape. Additional studies of this kind are needed, not only to replicate the findings in Verdu *et al.* (2017), but also to test for the presence of Wolof genotype among the Cabo-Verdean

population, which would provide further confirmation of the tenants of the Santiago-birth hypothesis.

In this section, linguistic evidence of the lexical, morphosyntactic, and phonological variety, were linked with language-external historical data, such as those concerning population movements, settlement patterns, and trade routes, in an early instantiation of the trans-Atlantic Slave Trade that implicated the Santiago, Senegambia, the UGC, and the Iberian Peninsula. With respect to linguistic evidence, the overwhelming correspondences across varieties of UGCP in the domains of lexicon and morphosyntax (both in the 'surface' shape of lexemes and their morphosyntactic function) make unlikely the notion that insular and mainland varieties of UGCP could have formed separately along parallel trajectories. Such extensive correspondences in so many domains of language could not have emerged unless there were a shared proto-vernacular source. This is especially true for the shared morphosyntactic features, which are argued to be absent from pidgin languages during the earliest formative stages of contact-driven linguistic restructuring (Jacobs 2010).

The presence of Classical Portuguese lexemes (Jacobs & Quint 2016), which had fallen out of use by the 16th-17th century, but are nonetheless present in UGCP, suggest that proto-UGCP must have been well on its way to formation/consolidation in the 15th century when Santiago was being settled. The lexemes in question could not possibly have entered UGCP via a proto-vernacular on the mainland. As research in the fields of history and anthropology confirm, Luso-African society on the mainland did not really begin taking shape until the end of the 16th century when the first Luso-African settlements of the '*ria* coastline' region began to appear.

Further, extra-linguistic support for this notion comes from the sociohistorical circumstances that shaped the language ecology of the UGC during this period; the distribution of societal power, population dynamics and the slave-based economic model on Santiago were such that it can be described as an early prototype for the slave-based societies based in ultramarine European settlements where contact vernaculars are known the have emerged repeatedly over the colonial era. Conversely, the geopolitical environ on the UGC mainland was in no way consistent

with the sociohistorical circumstances under which dramatic contact-induced language change, restructuring, or new vernacular formation are known to occur, since Lusophone settlers of the region were obliged to adhere to the tenants of the 'landlord-stranger relationship' (Brooks 2003[1993]; Baleno 2007:150). Thus, Portuguese could not be described as having had superstrate status, since Portuguese speakers had not achieved sovereignty over the region, nor did they settle there in sufficient numbers to have been in the position to induce the kind of language change consistent with the degree of restructuring observable in UGCP. Meanwhile, it was noted that, rather than Portuguese settlers, the ranks of the *lançados/tangomãos* was comprised primarily of *santiaguenses*, who by the turn of the 17th century had already firmly established a Luso-African proto-UGCP-speaking culture on Santiago. This means that proto-UGCP speakers from Santiago were the 'founder population' for the Luso-African settlements of the UGC mainland.

Along similar lines, substrate features in UGCP were found to be mostly of Wolof and Mandinka origin. Historical evidence shows that both these languages were not spoken in the '*ria* coastline' region of the UGC in sufficient numbers to be consistent with their extensive impact on the shared African-origin lexicon across varieties of UGCP, or to have contributed the morphosyntactic reflexes of purported Wolof origin, if one assumes a continental birth hypothesis. However, it is known, again from extra-linguistic historical evidence, that there were robust trade ties and population movements between Santiago and the Senegambia region, particularly in the first century and a half after settlement of the former, during precisely the period when the proto-UGCP vernacular would have been taking shape. Population-genetic evidence confirms what the distribution of substrate features in UGCP indicate, and what the evidence from historical trade routes and population movements would suggest.

Lastly, with respect to phonological features in UGCP, sound changes affecting phonemes from Classical Portuguese were observed to have resulted in differential outcomes in insular and mainland varieties of UGCP. The most likely direction of these sound changes allow one to infer that the Santiago variety of UGCP retained some of these conservative features from Old Portuguese, while on the mainland those phonemes underwent sound changes that indicate a

transfer of proto-UGCP from Santiago to the mainland, but not *vice-versa*. In the next section, I return to the language-external history of Santiago and Cabo Verde, picking up where I left-off in the 17th century and continuing through the end of the colonial era in the 20th century.

2.4 THE HISTORY OF CABO VERDE: MID-17TH CENTURY THROUGH THE END OF THE TRANS-ATLANTIC SLAVE TRADE.

The period of decline that had begun in Santiago between the end of the 16th and the start of the 17th century would continue and intensify through the duration of the latter century. Within the broader context of the trans-Atlantic Slave Trade and the Atlantic world, Santiago's decline could be attributed to a shift in trading patterns away from Senegambia and the UGC and towards equatorial West-Central Africa including the so-called Gold Coast and Slave Coast, Dahomey, Kongo, São Tomé, and Angola (Lobban 1995:27; Green 2006:270-271). But Santiago's continued decline in the 17th century was also driven by local forces. It was spurred on by cycles of drought and famine, the constant threat of pillage and violence from pirates and competing European powers, and Santiago's near total loss of political, strategic, and economic relevance in the Atlantic sphere.

The *armadores/moradores* of Santiago had been losing control of the regional trade to the French, Dutch, and British traders, and even other competing Portuguese merchants, since the second half of the 16th century. Navigators and merchants were also increasingly willing to bypass Santiago, preferring the Europe-UGC/Senegambia–Americas trade route. To make matters worse, on Santiago, after several years of drought, the famine of 1609-1610 devastated the island and marked the start of the collapse of Santiago within the Atlantic market (Torrão 2007:162-169; Green 2006:192). The loss of Gorée to the Dutch in 1629 effectively ended the little commercial activity Santiago still commanded in Senegambia, while back on Santiago power struggles plagued the local administration, the illicit trade flourished unchecked, and among the *morador* class many were unable to continue to maintain the chattel system, resulting in large numbers of manumissions

(Green 2006:241-243). Baleno (2007b:154-156) describes how in the first quarter of the 16th century *moradores/armadores* and *fazendeiros* were abandoning Santiago for the UGC mainland and elsewhere, having given up on the economic viability of Santiago and having become unable to cope with constant droughts, famines, and sacking.

With the 1640 *restauração* 'restoration' of the Portuguese monarchy, the flow of American-bound ships calling port in Santiago slowed from a trickle to intermittent drip, and things were only made worse in 1644 when traders bound for the Americas were officially allowed to pay levies and taxes at the *feitoria* 'factory' in Cacheu rather than Santiago, and in particular those bound for the *engenhos* 'sugar cane mills' of Brazil (Torrão 2007:168-169,178; Baleno 2007c:337; Green 2006:270-271). The Jesuit mission in Santiago ended in 1646 after years of financial difficulty and lukewarm enthusiasm (Green 2006:273), and upon official orders, the *feitoria* in Santiago was closed in 1647 (Torrão 2007:179), eventually to be replaced by the boom-and-bust of the *companhia* 'trade company' system. By 1662 the Dutch had been awarded a Spanish *asiento* for the trafficking of enslaved people to the Americas (Rodney 1970:96; Green 2006:271), an event that formally ended the official Santiago-Spanish Americas trade, although the informal illicit trade would continue well beyond the 17th century (Green 2006:248). Baleno (2007c:341) describes the transition:

> O comércio informal e o contrabando foram as principais manifestações das mudanças ocorridas na economia do aquipélago para enfrentar a crise resultante do estrangulamento do circuito Cabo Verde/Guiné de meados de seiscentos até à instalação da companhia Grão Pará e Maranhão, em meados do século XVIII" [Informal commerce and contraband were the main manifestation of the changes that occurred in the archipelago's economy in order to combat the crisis that resulted from the restriction of the Cabo-Verdean/Guinea circuit from the middle of the 1600s until the installation of the Grand Pará and Maranhão Company, in the middle of the 18th century.]

The combined forces of drought, famine, inflation, and overall loss of economic and strategic relevance in the Atlantic sphere triggered, by the middle of the 17th century, a dramatic ruralization of the population of Santiago - that is - a growth of the population in the rural interior of the island and a shrinking of its urban population (Baleno 2007c:341-342; Correia e Silva

2007c:299-304; Green 2006:242-245). The project of the elite *mordaor/armador* class shifted from an outward-looking, sea-bound enterprise, inextricably tied to the trans-Atlantic Trade, to an inward-looking agricultural enterprise, focused on the production of goods for local consumption and the provisioning of ships, a process described by Correia e Silva (2007c:299) as a transformation from an *armador-proprietário* 'ship-owner/landowner' class into an *agro-proprietário* 'agrarian-landowner' class. Whereas before all economic activities served the purpose of facilitating trade with the mainland under the auspices of royal contracts, now the economy came to be driven by the so-called *comércio miúdo* 'little commerce' (~'informal economy'): animal husbandry, agriculture, and the manufacture of orchil dye and *pano di terra* 'cloths of the land' (Baleno 2007c:341-342). Meanwhile, periodic cycles of drought and famine drove larger and larger-scale manumissions, most of the formerly enslaved people fleeing to the mountainous interior and deepening the ruralization process (Correia e Silva 2007c:309), while many others fled to the *ria coastline* region of the UGC mainland to continue to seek their fortune in the Slave Trade. Baleno (2007c:341) describes the transformation:

> Desaparece a figura do morador-armador e surge a do atravessador, os navios de longo curso dão lugar às pequenas lanchas de cabotagem, os mercadores do reino são substituidos por estrangeiros, cessam os regimentos da alfândega e segue-se o << uso e costume>>" [The figure of the landowner-shipowner disappears and the figure of the small-time merchant broker appears, the long distance ships give way to small boats for cabotage, the merchants of the kingdom are substituted for foreigners, international customs regulations cease and are supplanted by <<habit and popular convention>>.]

Correia e Silva (2007c:304) provides a similarly effective summary of the 17th century economic crises:

> Algumas das suas manifestações mais exuberantes traduziram-se numa baixa de rendimentos, no crescimento da dívida, das instituções e dos particulares, na desmonetarização, na desescravização, na redução do número de brancos, no aumento dos escravos fugidos e da frequência das fomes, na multiplicação dos crimes contra a autoridade, etc." [Some of (the crisis') most profuse manifestations were rendered as reduction in revenues, the growth of debts, (both) institutional and personal, in de-monetization, in the unraveling of the slave system, a reduction in the number of whites, an increase in the number of fugitive slaves and the frequency of famines, in the multiplication of crimes against the authorities, etc.]

Baleno (2007c:361) concurs that:

> deu-se a interiorização e ruralização da sociedade e o empobrecimento e <<crioulização>> dos homens poderosos. Nessa nova conjuntura nasceu uma nova elite, virada para a terra sua única riqueza. [there initiated the interiorization and ruralization of the society and the impoverishment and <<creolization>> of the powerful men. Out of this new affair was born a new elite, oriented towards the land(,) their only remaining wealth.]

Green (2006:242, 248, 255-256, 272-278) views the economic transformation, ruralization, and marginalization of Cabo Verde in the Atlantic sphere as driving "the emergence of the Creole power block by the 1650s," following protracted conflict among rival social/racial groups in Santiago. A Jesuit's letter compiled in the *Monumenta Missionária Africana* (Brásio 1962:613) portrayed the social hierarchy of Santiago as consisting of four groups in descending order of overall share of the free population from most to least populous:

> crioulos, que são os naturais da terra, christãos nouos, clerigos da terra e de Portugal, e algũs christãos velhos de Portugal, mas muito poucos." [creoles, who are native born, new Christians, native-born clerics and (those) from Portugal, and some old Christians, but very few.]

As resources and political capital became scarcer, these groups increasingly vied for power. Cabral (2007b:360) and Green (2006:255-256) argue that this power shift, provoked by the ruralization process, allowed the more populous creole sector to fill the ranks of the *morador/armador* class, resulting in the 'Creole power block' becoming the dominant social group by mid-century. This turn of events was facilitated by the "pauperization" of the traditional landowning class (Correia e Silva 2007c:305), and the ability of the newfound creole elite to coerce the various institutions of power on Santiago: the *alfândega* 'customs house', *a casa da misericórdia* – an organization within the Church in charge of public social assistance and religious indoctrination – the judiciary, the *câmara municipal*, and the militia, the latter of which sometimes rallied multitudes of freed formerly-enslaved people and the peasant class of the interior (Cabral 2007b:360-362).

Despite the rise of the 'Creole power block', Portuguese authorities still maintained a military presence in Ribeira Grande, although "it was the Portuguese power that lived in the

shadow of Creole society, rather than vice versa" (Green 2006:278). Though the 17th century saw among the most suffering of any century in the history of Santiago, the decline of the externally-oriented economy did correspond to a certain revitalization of internal subsistence economy (Correia e Silva 2007a; Green 2006:275). Green (2006:280-281) also views this period as one of consolidation of Kriolu cultural and linguistic identity, driven by the ascendance of the 'Creole power block'.

Long before the crises of the 17th century, the *arrendamento* system of royal contractual leasing of trading rights between Santiago and the mainland had already proved inefficient and outdated. Contract holders periodically failed to uphold the terms of the contracts and accumulated massive debts resulting in revocation, creating massive inefficiencies in the procurement, movement, and taxation of goods, and putting the Portuguese at a disadvantage relative to other European powers that were booming economically from the use of trade-company system (Baleno 2007c:349-350). The last quarter of the 17th century, and the better part of the 18th, would see many failed attempts at the establishment of *companhias* 'trade companies' that aimed to maximize the profits extracted from the trade in orchil, *pano*, and enslaved people from Cabo Verde/Guiné to Brazil. This was a market which, as we have seen, had now been reduced almost completely to unsanctioned trade that evaded the scrutiny of the royal administration; the *companhias* proposed to restore the officially-sanctioned management and profiteering apparatus, this time under the auspices of a modern, capitalist enterprise rather than the feudal contracts of earlier centuries, but with the same 'full' accounting of revenues destined for the Crown

The *Companhia de Cacheu, rios e comêrcio da Guiné* (CCG), founded in 1675 by investors from Lisbon and run out of Bissau, was awarded a 6-year contract for the shipment of enslaved people. Some of them were destined for Santiago, but most were destined for the metropole, from where they would later be sent to the Spanish Americas or other parts of the ultramarine Portuguese Empire (Lobban 1995:27). After the establishment of the CCG, the contractual rights of *armadores* from Santiago to engage in trade with the mainland remained unchanged. However, as we saw above, the *morador/armador* class had essentially lost their ability to rig ships and conduct trade

operations with the mainland. This situation put the *moradores/armadores* of Santiago at the mercy of CCG, whose monopoly status over the external trade allowed them to purchase goods from Santiago at artificially low prices, sell foreign goods at artificially high prices, and otherwise dictate the terms of trade (Baleno 2007c:350).

The grip of the CCG was reinforced by legislative measures in Lisbon, such as a 1680 *alvará* that insisted only metropolitan merchants could ship *pano* (Correia e Silva 2007c:309). Ultimately, however, the operations of CCG alienated the *lançados* operating in the Rios da Guiné, who continued to resist any efforts on the part of the Crown or related agencies to levy taxes or otherwise limit their freedom of movement and commerce. As we have seen, the *lançados/tangomãos* had long since become masters of the illicit trade and openly did business with merchants and companies of other European powers in competition with the Portuguese Empire (Lobban 1995:27). These themes, with the *moradores/armadores* of Santiago at the mercy of the *companhia,* and the *lançados/tangomãos* of the mainland subverting the efforts of the *companhia*, would be repeated over the course of the *companhia* era.

Aggressive challenges from other European powers continued in the forms of raids, sackings, and general piracy. In the worst attacks since Francis Drake's raid in 1585, French corsairs attacked and briefly held Ribeira Grande and Praia in 1712, prompting many to flee to the interior and causing extensive damage to goods and property (Carreira 1972:345; Lobban 1995:28). Meanwhile, both the French and English were becoming more dominant in trade with the UGC mainland, and famines became more regular occurrence on Santiago. Correia e Silva (2007c:309) lists records of droughts and famines in: 1660-1662, 1685-1689, 1704-1712, 1719-1721, 1730-1732, 1738-1740, 1741-1742, 1745-1746, 1748-1750, 1754-1755, 1764-1765 and 1773-1775. As a result, the ruralization process and the shift in the economy from an external to an internal orientation, continued well into the 18th century.

King José I, who reigned from 1750-1755, along with his secretary of state the Marquês de Pombal, oversaw the most impactful policy changes for Cabo Verde in the 18th century. Together, they embarked on a program of reforming feudal privileges, modernization, centralization, and

secularization of the imperial bureaucracy. Among these were included the establishment of an imperial police force, a royal treasury, a university system, and royal printing press, the abolishment of slavery in metropolitan Portugal, and after the attempted regicide of José I in the 1758 *Processo dos Távoras*, the expulsion of the Jesuits and the imposition of major limitations on the power of the nobility (Lobban 1995:29).

On Santiago these reforms meant that the remaining *donatários* 'land tenure contacts' were sold, the capital was transferred from Ribeira Grande to Praia, and there was a strengthening of the administration's direct authority on local affairs (Lobban 1995:29; Cohen 2007b:317). Another reform involved the foundation of a new trading company, the *Companhia Geral do Grão Pará e Maranhão* (CGGPM), which was intended to revitalize trade with the UGC while encouraging development in the sparsely populated Brazilian northeast where the British had Dutch had been challenging Portugal's sovereign claims for years. The CGGPM was granted a monopoly over the trade in *pano*, orchil, and enslaved people, with their center of operations based in Santiago, and their primary mainland trading post at Bolama. It was notorious for its brutal treatment of enslaved people, and its extractive impact was unprecedented, shipping a total of 28,167 enslaved people from either Guiné or Cabo Verde to Pará and Maranhão between 1756 and 1788 (Lobban 1995:30).

Though the CGGPM earned massive profits for investors in Brazil and Portugal, Santiago saw little to no economic improvement, since as with the CCG, the CGGPM was able to dictate the terms of the trade and left the *moradores* "à mercê da companhia" [at the mercy of the company.] Baleno (2002:350) describes how:

> a companhia faz tudo para dificultar a actividade dos moradores. Quando estes conseguem resgatar alguns escravos, deparam-se com a questão do transporte, porque, segundo afirmam, em Cacheu havia <<ordens para que não recebessem escravos dos passageiros ou mordadores desta Ilha>>, no intuito os forçarem a vender <<as ditas peças a mesma Companhia pelos preços que muitas vezes lhes não acomoda >> [the company does everything they can to impede the activity of the landed elites. When they manage to purchase some slaves, they encounter the issue of transportation, because, according to what is claimed, in Cacheu there are <<orders for slaves coming from the Island not to be received>>, with the intention of forcing them to sell <<said cargo to the Company itself at prices that often are not convenient for them.]

The grip the CGGPM held over Cabo Verde was indeed extensive:

> pela abrangência dos privilégios solicitados, assiste-se a um processo de privatização do arquipelago, substituindo-se a companhia ao Estado nas suas atribuções de soberania [due to the scope of the privileges that had been ordered, one sees a process of privatization of the archipelago, substituting the company for the State in their applications of sovereignty] (Baleno 2007c:352-353).

As we can see, the reforms of the Marquês de Pombal generally worked against the interests of local elites in Santiago, and in 1764 metropolitan Portuguese troops were installed on the island to enforce Lisbon's mandates, while in 1770 the powers of *governador* were extended to check the power of the *câmara* 'municipal chamber' and the *ouvidor* 'councilor', who had supported the interests of the 'Creole power block' (Cabral 2002:374).

Just as was the case with the CCG, the extractive and exploitative power of the CGGPM catalyzed the pauperization of the landholding class, the ruralization Santiago, and the growth of the population in its interior, mostly driven by manumission and maroonage (Correia e Silva 2007c:309-310). Moreover, the process of ruralization had contributed to an acute labor shortage by the middle of the 18th century, which left landowners unable to extract profits from their landholdings and forced them to parcel out their *latifundios* 'estates', creating a self-reinforcing cycle (Carreira 1972:388). The confluence of the land crisis, the already harsh climactic conditions on Santiago, and the extractive, abusive, and neglectful policies of the CGGPM, triggered the brutal famine of 1773-1775, which left as much as 40% of the population to die of starvation (Bigman 1993:80). Finally, in 1778, the CGGPM ceased operations, ultimately brought down by corruption, over-extension of resources, and resistance from *lançados* operating on the mainland (Lobban 1995:30).

The 19th century, as the previous two centuries in Cabo Verde, was defined by yet more cycles of drought and famine, an increase in the number of exiled *degredados* sent to the island from metropolitan Portugal and elsewhere in the empire, rebellion and uprising by various marginalized groups including enslaved people, *degredados*, and underpaid troops, and by the desperate attempts of the Cabo-Verdean merchant class elites to hold-on to the vestiges of the

Slave Trade, even as the rest of the imperial world was phasing out the practice (Carreira 1972:362, 1982:25-26).

Ribeira Grande had now fallen into complete decay, and the ruralization process of 17th and 18th centuries had advanced far enough that "one ought not to speak of an urban bourgeoisie" (Carreira 1982:25). While Santiago and Cabo Verde remained in stagnation, the broader Atlantic sphere was entering a period of upheaval in which the trans-Atlantic Slave Trade model was being challenged by abolitionist movements throughout the Americas, coupled with abolitionists movements in the primary slave-trading nations of Europe.

The Slave Trade Act of 1807 in the United Kingdom marked the begging of the period in which the UK would start to pressure other major European imperial powers to cease their participation in the trans-Atlantic Slave Trade. An 1815 British-Portuguese treaty signed in Rio de Janeiro prohibited the Portuguese Slave Trade on the West African coast north of equator, other than for those enslaved people destined for Portuguese possessions; that is, ships sailing under the Portuguese flag were not to trade in enslaved people with other nations in the Atlantic sphere, but they could send enslaved people from one Portuguese port to another (Carreira 1972:395).

The Spanish Slave Trade also began to feel the pressures of British abolitionist efforts, and in response, rather than lessening their participation in the trade, they doubled-down and intensified the volume and frequency of the trade, particularly for those enslaved peoples destined for Santo Domingo and Cuba (Carreira 1972:395). Despite British pressure on the Spanish and Portuguese, a covert Slave Trade persisted on quasi-legal grounds, there being evidence of Spanish firms contracting and provisioning ships in Santiago to sail and trade in enslaved people under the Portuguese flag. Some vessels, having been licensed and provisioned in Santiago, and thus in compliance with treaty requirements, were able to skirt the authority of the British West African Squadron, despite the fact that their ultimate destination was often Spanish claims like Santo Domingo or Cuba (Carreira 1972:396-398).

Despite this last push on the part of Spanish and Portuguese merchants, the chattel slavery system on Santiago itself had long since been receding. As Lobban (1995:31) explains:

> Only a few hundred of the wealthier and landed families, absentee landowners, and nobility had sufficient means to afford such unsalaried servants and laborers. […] Of those who owned slaves, about a third had only one or two to perform domestic chores. […] If one includes all slaves held in Cape Verde at this time, the average slave staff was three or four. Although some gangs of slaves did toil in the fields and mines, rarely in Cape Verde did a single proprietor hold twenty or thirty slaves. In fact, the largest island plantation was the only place where forty or sixty slaves worked on a single farm. The colonial state held slaves for labor on public works and in construction, and the Catholic Church owned slaves for cultivating capela lands and maintaining church buildings.

The Marquês Sá da Bandeira, as Prime Minister of Portugal in 1836, oversaw a declaration abolishing trans-Atlantic Slave Trade under the Portuguese flag, although the statute included exceptions for ships sailing from Guiné to Santiago that were allowed to ship up to 10 enslaved people at a time (Carreira 1972:401). Another Anglo-Portuguese treaty in 1842 prohibited the importation of enslaved people to Santiago and increased the authority exercised by the West Africa Squadron in the region (Carreira 1972:398,407; 1982:27). An 1856 census of enslaved people in Cabo Verde, and elsewhere in the ultramarine Portuguese Empire, was intended to tally the soon-to-be-freed population, the end of state-sanctioned slavery being just on the horizon (Carreira 1972:416). This census revealed there to be 5,180 enslaved people in Cabo Verde and a total of 1,194 on Santiago (Lobban 1995:31-33).

In 1858, a 20-year deadline was set for the total abolition of slavery, and then in 1875 this deadline was reset for the next year, 1876, after which enslaved people were obliged to sign contracts, with preference going to the previous slave-owner, for a period of two years of indentured servitude (Carreira 1972:416). Clearly, as slavery was being phased out, a new type of indentured servitude-like 'wage-slavery' would lock peasant workers into toiling under life-threating working conditions for menial pay.

Beginning in 1863 and continuing into the 20th century, the victims in Cabo Verde of droughts, crop shortages, and famine, were summarily sent to the island of São Tomé in the Gulf of Guinea to toil on the cocoa plantations. This was a top-down response to the labor shortage following the *de facto* abolition of slavery on that island (Bigman 1993:85). Often Cabo-Verdeans sent to work in São Tomé found themselves unable to escape a system of debt and starvation wages

that saw them unable to return to Cabo Verde. The harsh conditions on the plantations of São Tomé meant few survived beyond their 40s. As many as 24% of those who went to São Tomé between 1911-1928 passed away there.

Tragically, the plight of so many Cabo-Verdeans in the first half of the 20th century was scarcely better than under chattel-slavery. The second half of the century, however, would bring newfound opportunities for Cabo Verdeans to win their right to self-determination and self-governance, finally throwing off the yoke of Portuguese imperialism. In the next section, I summarize the history of the final stages of the colonial period in Cabo Verdean and the path to independence.

2.5 POST-SLAVERY SANTIAGO, THE MODERN COLONIAL ERA, AND INDEPENDENCE.

Unlike so many of their colonial counterparts in the Americas, the African Portuguese colonial possessions remained firmly under imperial control when European ultramarine territorial claims were formalized at the Conference of Berlin in 1884–1885. The formerly enslaved people of Cabo Verde now became a peasant class locked into a system of indentured-servitude-like wage-slavery. They toiled as *parceiros* 'share-croppers', *rendeiros* 'tenant farmers', *contratados* 'contract laborers', *brigadas de estrada* 'road work gangs', or *frentes de trabalho* 'unskilled labor force' (Lobban 1995:41). The formal abolition of the *morgadio* system did little to slow the mass parceling-out of land by landlords to the *parceiros* and *rendeiros*, exacerbating the problems of scarce arable land, over grazing, and soil depletion, all of which worsened the drought-famine cycle (Carreira 1982:28-29).

Carreira's (1982: 29) translation of a description of the flawed post-slavery system that kept Santiago mired in poverty written by *governador* João Cesário de Lacerda in 1898 is telling:

> Many owners of vast rural properties which could be cultivated on a large scale to provide abundant export crops, are content to let them out in small patches for rent money or a share of the crop. They are generally let out for a year at a time, and the tenants confine themselves to sowing their lands with enough to provide them with foodstuffs for their ordinary use. The tenant tries to obtain nothing more, apart from

> setting aside what has to be paid in rent, since he is neither inspired by incentives nor driven by necessity. This system has many disadvantages and is one reason why the economy of the province is so backward. […] Any tenant who improved his land and introduced long-term, extensive cultivation, could be sure that at the end of the year the proprietor would only consent to renew the lease in return for an increase in rent, so that it would be only he who derived any benefit from the improvements.

The hope for liberal change on Santiago that was promised with the 1910 coup d'état and the installation of the Portuguese Republic was quickly quashed by the rise of dictator António Salazar and his *Estado Novo* regime in 1926 (Rego 2015:42). In metropolitan Portugal, the cause of liberation for the African Portuguese colonies, driven primarily by students from across Lusophone Africa attending universities in Portugal, had become closely associated with the activities of the Communist Party, which encountered staunch repression from the *Polícia Internacional e de Defesa do Estado* (PIDE) (Lobban 1995:43).

In Cabo Verde, the emerging liberation movement and cultural revolution of 1930s was iconically expressed in the famous literary journal *Claridade*; based out of Mindelo, São Vicente, it counted among its authors and intellectuals such as Jorge Barbosa, Manuel Lopes, and Baltasar Lopes da Silva, the latter of whom published a rigorous linguistic description of Kriolu (1959) and perhaps the most famous Cape Verdean novel, *Chiquinho*. *Claridade,* and the movement associated with it, embraced creole identity, dealt with local issues such a colonial neglect, drought, famine, and abject poverty, and offered up subtle critiques of the fascist regime and Cabo Verde's colonial plight, all while bridging the gap between the Romanticist and Realist eras in Lusophone literature (Passos 2003; Rego 2015:43). The movement soon encountered censorship and surveillance from PIDE, and in 1938 the infamous concentration camp in Tarrafal, Santiago opened. This detention center would come to be synonymous with Salazar's brutally repressive state and imprisoned hundreds of political prisoners sent from Portugal and all-over Lusophone Africa.

State repression was hardly necessary, however, since economic woes, colonial neglect, and the drought-famine cycles kept Cabo-Verdeans in a state of poverty and hunger, alas, too weak or preoccupied with survival to muster anything like a resistance movement. In the first half of the

20th century, Cabo Verde experienced at least seven major famines and droughts, all of which are well documented and compiled in Carreira (1984:43-109); he provides an estimate of as many as 50,000 famine related deaths in Cabo Verde over the first half of the 20th century. Condemnation to work on the cocoa plantation of São Tomé continued to be the administrative response to famine, and during the first half of the 20th century another 50,000 Cabo Verdeans met this tragic fate (Bigman 1993:85).

In the face of mass suffering, by the 1950s, revolutionary independence movements had emerged across Lusophone Africa. Within the context of the Cold War, they were often associated with Marxist and Maoist ideologies of popular urban working-class struggle and rural peasant revolt; in reality, for the Luso-African revolutionary movements, national liberation was always prioritized over ideological pursuits and alignments with the Soviet Bloc states was usually viewed as a matter of necessity considering Portugal's membership in the NATO alliance (Lobban 1995:88). Additionally, nascent liberation movements across Lusophone Africa often had direct ties to the various leftist movements of Portugal, since those from the Lusophone African nations in the 20th century who pursued higher education were often obliged attend universities in the Portugal where they became involved with liberal-republican, socialist, and communist underground student organizations operating just beyond the scrutiny of the Salazar regime (Rego 2015:45).

In *Guiné*, social unrest and violent conflict between the state and many ethnolinguistic groups including the Papel, Fula, Balanta, and Bissagós had continued on-and-off for the first half of 20th century, but these had been localized affairs, lacking the bonds of a national or international unity movement. Now in the 1950s, cultural ties and personal ties between Cabo Verde and *Guiné* were being revitalized within the context of pan-African Liberation. In 1956 the *Partido Africano para a Indepenência da Guiné e Cabo Verde* (PAIGCV) was formed by intellectuals from the islands and mainland, such as brothers Amílcar and Luís Cabral, Aristides Pereira, and Rafael Paula Barbosa, among others (Lobban 1995:44,88).

The first major event for the PAIGCV after years of organizing and forging alliances with international Marxist and African Liberation movements was the Pijiguiti Dockyards Massacre of 1959. Police forces had become aware of the alliance between the striking dock workers and the PAIGCV and opened fire on the striking workers, killing around 50 and wounding over 100 (Lobban 1995:89-90). In response, most of the members of the PAIGCV fled into newly independent Guinea-Conakry to the south, and the group took the opportunity to regroup and reassess their liberation strategy.

It was during this time that the PAIGCV dedicated themselves to an armed, rather than peaceful liberation struggle, their target for recruitment became the rural peasantry, and combat became focused on the rural, rather than the urban theater (Lobban 1995:90). Senior PAIGCV members also continued to seek diplomatic alliances internationally, the 1961 *Conferência das Organizações Nacionalistas das Colónias Portuguesas* held in Casablanca saw the incorporation of other Cabo Verdean and Guinean revolutionary movements into the PAIGCV, and reaffirmed unity and collaboration with the revolutionary forces of other Lusophone African nations such as the *Movimento Popular de Libertação de Angola*, the *Frente de Libertação de Moçambique*, and the *Movimento de Libertação de São Tomé e Príncipe* (Lobban 1995:91).

In 1963 the long planning and diplomacy phase ended, and the armed conflict began. The PAIGCV opened a front in southern *Guiné-Bissau* near the border with Guinea-Conakry. Using guerrilla tactics, they quickly came to control substantial territory south of the Corubal River (Lobban 1995:92). The next year the party held its first congress in which the it expanded its bureaucracy and began to acquire the apparatus of a formal state, while local strongmen who had abused their power were summarily removed from the party in an attempt to maintain good relations with the rural peasantry (Lobban 1995:92-93).

In response to the advances of the PAIGCV, the Portuguese increased the troop presence in *Guiné* to 12,000–14,000 and continued to hold the more populous western coastal regions, while the majority Muslim Fulas in the east had not yet been persuaded to join the PAIGCV cause. Despite this, by 1965 the PAIGCV controlled approximately half of Guiné-Bissau and continued

making territorial gains while building their case for liberation on the international diplomatic circuit (Lobban 1995:93-95). Becoming desperate, under the supervision of General António Salazar, Portuguese special forces in collaboration with Guinean resistance fighters carried out *Operação Mar Verde*, an amphibious assault on Conakry aimed at striking and capturing the PAIGCV members harbored there, liberating Portuguese POWs, destroying ships and aircraft, and deposing the democratically elected government of Ahmed Sékou Touré (MacQueen 1997). This operation failed to assassinate or capture PAIGCV leaders or Touré, and the Portuguese Special Forces retreated having only freed the POWs and destroyed some military assets.

Amílcar Cabral, the famed PAIGCV leader born in *Guiné* to Cabo-Verdean parents, led diplomatic efforts for the PAIGCV on the international stage. He lobbied the UN for recognition of Guiné-Bissau in the General Assembly, which in 1972 sent a mission to the PAIGCV controlled areas. That same year the PAIGCV held party elections in the occupied territories, planned for the first elections of the *Assamblea Nacional* in 1973, and held an independence referendum (Lobban 1995:96-97).

In 1973 Amílcar Cabral was assassinated in Conakry by Inocêncio Kani, a PAIGCV member who according to some accounts was influenced by Portuguese PIDE agitators, or fell victim to ethnic tensions, or tensions within a wing of PAIGCV that had been assimilated from another militant group (Lobban 1995:98). Sékou Touré then halted an attempt to arrest and take to Bissau the remaining PAIGCV leadership, and later as many as 100 PAIGCV members suspected of having participated in the plot were allegedly summarily executed (Lobban 1995:98).

Back on the battlefield in *Guiné-Bissau*, the PAIGCV offensive could not be slowed, even after the execution of their revolutionary leader. Using newly acquired artillery and anti-aircraft weapons, PAIGCV forces were able to topple the fortified garrison at Guiledge, and that same year the second PAIGCV elections proceeded as planned. Important bureaucratic positions were elected and preparations for a new constitution were put into place (Lobban 1995:98-99). Similar gains for the other liberation struggles of Lusophone Africa also obtained, and with the compounding losses and mounting pressure from the international community, Portuguese General Spínola was

forced to step down in November of 1973. Finally, in 1974 the bloodless coup and the following social upheaval know as *a Revolução dos Cravos* 'the Carnation Revolution', led to the deposition of Marcello Caetano and the *Estado Novo* (Mailer 2012:103-113).

Though the stage seemed to be set for the inevitable independence of the Lusophone African nations, a resolution to the Cabo Verde and *Guiné* conflict was still in question. For one thing, virtually the entire revolutionary war had been carried out on the mainland, and though many Cabo Verdean were involved with the PAIGCV, no operations were being conducted on the islands until 1974 when PAIGCV began peaceful organization efforts there (Lobban 1995:100-101). These culminated in large rallies in Praia and Mindelo where Portuguese police opened fire on protestors (Lobban 1995:101). By August of that same year, arrangements were made for the granting of independence to *Guiné* in the Algiers Accords, while plans were being made to push for an independence referendum in Cabo Verde (Lobban 1995:102-103).

In the first elections, General Spínola became president of the Portuguese Republic and was still firmly in opposition to Cabo-Verdean independence, hoping instead to retain the archipelago as autonomous regions like *Açores* or *Madeira* (Lobban 1995:102-103). Independence rallies and workers' strikes were becoming regular occurrences and tensions were escalating between Portuguese military police and protestors, until a general strike on September 27th and 28th of 1974 brought tensions to a boiling point (Lobban 1995:102-103). In Lisbon, Spínola's own *Movimento das Forças Armadas* (MFA) - the military junta overseeing the transition to a republic - and *a Junta da Salvação Nacional* - the political wing of the *Comando Operacional do Continente* - all turned on Spínola and forced him to step down from the presidency in October of that year (Lobban 1995:105). The next year, the Portuguese Constituent Assembly was charged with drafting a constitution; Spínola attempted another coup but was unsuccessful and forced to flee to Spain and later Brazil (Mailer 2012:113-115).

Finally, the transition in Portugal to civilian rule was allowed proceed, if not slowly and chaotically. In June of 1975 election were held in Cabo Verde for a National Assembly in which about 85% of the population participated and the PAIGCV received 92% of the votes (Lobban

1995:108). On July 5th the independent Republic of Cabo Verde was declared, and the National Assembly met officially for the first time as the parliament of a sovereign nation (Lobban 1995:108). A new constitution was promised within 90 days, although it would not actually be forthcoming until 1980 (Lobban 1995:106-113).

The dynamic of post-independence was that PAIGCV was the ruling party both on the archipelago and the mainland, though each region had their respective National Assemblies. Meanwhile, a number of opposition forces, both on the archipelago and the mainland, were vying to play a role in the nascent Republic (Lobban 1995:107-112). In 1980 the first President of Guiné-Bissau, founding member of PAIGCV, and half-brother of the deceased Amílcar Cabral, was deposed in a bloodless coup d'état led by João Bernardo 'Nino' Vieira (Lobban 1995:112). Aristides Pereira, also founding members of PAIGCV, Secretary-General for the entire party, and simultaneously the first president of Cabo Verde, makes the decision to disjoin the party from its mainland counterpart (Lobban 1995:113). Henceforth, Pereira's party was to rule solely in Cabo Verde as the PAICV, completing the separation of the once unified republics (Munslow 1981). The PAIGCV would maintain single party rule until 1990 when it agreed to hold open elections. The nascent *Movimento para a Democracia* led by Carlos Veiga won a 2/3 majority in the *Assamblea Nacional* and succeeded in implementing constitutional reforms, formally allowing for a pluralistic representation (Lobban 1995:116-117).

In the next section, I provide a contemporary sociolinguistic sketch of Cabo Verde with an emphasis on Santiago. Included in this section is a discussion of the contemporary economic status of Cabo Verde, a brief overview of the debate over the officialization of Kriolu, issues of identity and attitude regarding linguistic choice and use, and some examples of linguistic variation conditioned by social factors such as long-term diglossia.

2.6 CONTEMPORARY CABO VERDE AND A MODERN SOCIOLINGUISTIC PROFILE OF CABO-VERDEAN CREOLE.

Since the opening of Cabo-Verdean politics to two-party rule, the young Republic has experienced no major civil strife, and has enjoyed gradual, moderate, but significant economic growth, and substantial improvements in quality of life for its citizens. Compared to the mass starvation, poverty, and brutal oppression of the colonial era, the improvement in the quality of life for Cabo Verdeans in the last quarter of the 20th century and first two decades of the 21st has been nothing short of remarkable, despite the on-going struggles with hunger, poverty, and limited public resources. Still, substantial improvements are yet to be made, particularly with respect to services provided by the state in the sectors of public sanitation, housing, education, and in access to alimentary and water resources.

Amado (2015:105) attributes much of the ongoing shortcomings in public services to limited participation in electoral politics, and virtually no participation in other sorts of civil institutions (both of which may be due to the power dynamics of diglossic societies):

> In spite of the discourse of democracy and the institutional innovations brought by the implementation of liberal democracy in the country, political participation is limited to electoral participation. Other forms of political participation either are non-existent or insignificant. In two decades of liberal democracy, citizens rarely challenged the policies, and/or the rules decreed by the government. A culture of political passivity still exists among a significant portion of the population. Collective actions, particularly in the form of peaceful political protest, are infrequent, sporadic, and characterized with low turnout. Moreover, other forms of political participation, chiefly those that require far more resources and time, such as contacting, in written format, the representatives or the state officials, are nonexistent. Other forms of participation such as taking part in public debate and discussion are limited by the institutional constraints. At the same time, mechanisms of open political debate such as public hearings on a legislative bill or political consultations are almost inexistent in the country. Other instruments of active and engaging citizenship such as initiative of law, while mandated by the Constitution, are yet to be implemented.

Much of the growth and prosperity enjoyed by Cabo-Verdeans since independence owes to the contributions of expatriate Cabo-Verdeans in the diaspora, as well as the growth of the tourism and shipping sectors. As we have seen, Cabo Verde has always depended greatly on its connections with the outside world, and the diaspora had been an indispensable feature of the Cabo

Verdean economy and cultural experience throughout the in the modern era. The largest Cabo-Verdean communities can be found in Southern New England in the United States, Portugal, France, Holland, Luxembourg, and São Tomé & Príncipe.

One of the earliest and most enduring diasporic communities can be found in the states of Rhode Island and Massachusetts, where connections from the trans-Atlantic Slave Trade linked Cabo-Verdeans slave-traders with major North American trade routes. Staring in the early 19th century, the whaling industry also brought Cabo-Verdeans to the shores of New England. The New England diaspora continued to grow in the first two decades of the 20th century with at least 20,000 mostly male Cabo-Verdeans traveling there. Although this migration was halted by major anti-immigration legislation starting in the 1930s, it commenced again in the 1960s, and today the community consists of approximately 300,000, making it the largest Cabo-Verdean diasporic community and one that maintains strong ties with the homeland (Batalha & Carling 2008:20; Quint 2009c).

As was mentioned prior, another longstanding Cabo-Verdean community is that of São Tomé, where Cabo-Verdeans were sent to work the cocoa plantations, usually under duress of force, starting in the mid-19th century, and continuing through the remainder of the colonial era. An estimated 80,000 Cabo-Verdeans settled in the equatorial archipelago during this period (Carreira 1983:245; Batalha & Carling 2008:21).

The 1960s and 1970s saw major migrations to Western Europe, and though there had been Cabo-Verdean communities in Portugal and the Netherlands virtually since the initial settlement of Cabo Verde, growth in the construction and shipping industries triggered larger-scale migrations, and today the Cabo-Verdean diaspora in Portugal numbers about 100,000, while Holland and France each have communities of around 50,000 (Batalha & Carling 2008:22; Quint 2009c). Other smaller diasporic communities exist in Argentina, Brazil, Spain, Italy, Luxembourg, Senegal, and Angola.

Culturally, the diaspora is hugely influential on the archipelago. Batalha & Carling (2008:15) described how the diasporic links to Portugal maintain strong cultural connections between Cabo Verde and greater *lusofonia*:

> The recent revival of economic ties with Portugal adds to a sustained Portuguese influence in many spheres of Cape Verdean society since independence. The institutions of the state and civil society are decidedly inspired by Portuguese models. In everyday life in Cape Verde, perhaps the most obvious example of Portuguese presence is football. Almost every man, and many women, in Cape Verde is a staunch supporter of one of the three main football clubs in Portugal: Benfica, Porto, or Sporting. Portuguese television shows and Brazilian soaps (novelas) are also important cultural influences from other parts of the Lusophone world.

Since the constitutional reforms of the early 1990s established six parliamentary seats for representation of the diaspora, and established an agency to deal with the affairs of expatriates, the diasporic vote has played a major role in the politics of the archipelago, even breaking a near tie in the 2006 elections where the diasporic vote tipped the scales in favor of the PAICV candidate Pedro Pires (Batalhtha & Carling 2008:22-25). Since the 2000s, the Cabo-Verdean economy has been undergoing somewhat of a transition, with revenues in expatriate remittances being surpassed by revenue from tourism (Batalha & Carling 2008:26).

The status of Kriolu in expatriate communities is not known with precision, though Quint (2005, 2012) estimates there to be a total of one million speakers of Kriolu, with around 450,000 in Cabo Verde, of whom 50% are speakers of Badiu. Around 250,000 live in the USA, 100,000 in Portugal, 50,000 in France, 50,000 in Holland, and smaller communities in the tens of thousands are in Senegal, in the thousands in Luxembourg and other Western European nations such as Spain and Italy, and in Angola. Rego (2015:85), following a consultation of the Ethnologue website (Lewis, Simons, & Fennig eds.) in 2014, places the total number of CVC speakers worldwide at 1,200,000, of which presumably less than half are living in Cabo Verde. My own consultation of Ethnologue in April of 2018 shows 492,000 speakers from the 2010 census, 80,000 of which were monolingual speakers, 325,000 of which were speakers of a Sotavento variety, and 167,000 were speakers of a Barlovento variety. The ethnic diaspora was listed at 1,000,000, the total number of

Kriolu speakers worldwide was listed at 734,130, leaving 242,130 diaspora speakers out of 508,000 people considered to be members of the diaspora[23].

Despite the uncertainty surrounding language use and fluency in the diaspora, it has always been true that the sociolinguistic status of Kriolu lies at an intersection of a trans-Atlantic crossroads. Historically, this complex relationship has been fluid, as the relative isolation of the archipelago has shifted in response to various forces global and domestic. Duarte (1998) and Coonan (2007:22), point out that during the era of the trans-Atlantic Slave Trade, there was no diglossia in Cabo Verde, since "[Kriolu] did not function as a dominated language [for several centuries], because it was practically the only language spoken, as much by Blacks and mestizos as by Whites" (Duarte 1998:159).

It was during the post-slavery colonial era and during the *Estado Novo* regime that a sharply contrastive diglossia emerged. Under official imperial policy, Kriolu was essentially ignored and was swiftly removed or repressed in any instance where it would emerge in the public sphere. Amado (2015:114) describes the policy as "glottophagic, that is, the systematic and calculated removal of the vernaculars from the colonial public sphere." Beyond the suppression of Kriolu wherever possible, public schooling was virtually absent on Cabo Verde, and since only the local elite could afford to go to Portugal to study, they alone could enjoy sufficient access to European Portuguese so as to acquire the language to an extent that the social privileges afforded fluent speakers of the metropolitan variety would be attainable (Amado 2015:133).

Literacy and public schooling have improved immensely post-independence. At an overall rate of 84.9%, Cabo Verde ranks in the top ten in Africa in literacy (UNESCO 2019). These figures *de facto* represent literacy in Portuguese, since CVC is still not taught in schools, and the remaining 15% percent who are illiterate come almost completely from older generations who were of school-age during or immediately following the colonial era (Oquendo 2014). So, while public education

[23] It is unclear to me exactly how these numbers for Kriolu-speaking in the diaspora were calculated by Ethnologue.

has improved immensely, with respect to linguistic power relations, post-colonial institutions have changed little. As Amado (2015: 115) explains:

> Many scholars who study post-colonial politics and society in Africa agree that there is an institutional continuity between the colonial and the post-colonial states […] The process of institutional continuity is particularly observable in language educational policy, that is, the choice of the language to be used as the medium for schooling.

This is evident in the PAIGCV's decision to maintain Portuguese as the sole language of instruction despite popular enthusiasm at the prospect of elevating the status of CVC (Oliveira Almada 1998:128; Amado 2015:144).

The debate over the status of CVC has been alive and well among Cabo-Verdean intellectual elites, best evidenced by the 1979 Mindelo colloquium led by Dulce Almada Duarte. This conference resulted in the first agreed-upon orthographic standard for CVC. These followed a mostly "phonetic-phonologically" motivated system, rather than an etymologically-based one that reflected more closely Portuguese orthographic norms (Table 11) (Coonan 2007: 119-120; see also: Baptista 1997; Coonan 2007:110-111; Amado 2015:147-148).

Table 11. Graphemes in Portuguese and two alphabets for Cabo-Verdean Creole and their corresponding phonemes (from Coonan 2007:119–120)

Phoneme	Graphemes		
	Mindelo Alphabet	ALUPEC	Portuguese
/ʃ/	ŝ	x	ch, s
/tʃ/	ĉ	tx	ch
/dʒ/	ĵ	dj	d (*Brazil*)
/k/	k	k	c, qu
/s/	s	s	s, ss, c, ç
/z/	z	z	s, z

Following the opening of the multi-party system, a newly proposed CVC orthography began to be developed. It was eventually approved in 1998, and is known as ALUPEC (*Alfabeto*

Unificado para a Escrita do Caboverdiano 'Unified Alphabet for Writing Cabo-Verdean') (Baptista 1997; Amado 2015:148; Rego 2015:54 – 55). ALUPEC is now the preferred system for writing in CVC in most academic and journalistic settings, although in popular usage it seems to be mostly ignored. While the state has endorsed its use and validity as the standard CVC alphabet, they have yet to implement its practice in any substantial way, not in the communications and procedures of the state, nor in the education system, nor other public domains (Coonan 2007:91-92).

Popular usage of CVC in online social media, in so far as can be gleaned from casual non-scientific observation, appears to vary in the degree to which the author opts for etymological or phonetic/phonological-based norms, though informal norms consistent with the latter seem to be emerging; of course, this question is best left to empirical inquiry. Within academic, governmental, and other elite sectors, there continues to be embattled debate over how CVC should be represented orthographically; a debate which Coonan (2007:110) considers to be a proxy for the "[…] struggle over whether or not Kriolu can and should be made autonomous from the Portuguese language".

The challenge of representing CVC's rich regional variation is also a point of contention. Each island of the archipelago boasts its own autochthonous regional variety, or in the case of the larger islands like Santiago and São Vicente, several regional variants. Generally, though, varieties of CVC are grouped together according the traditional geographic groupings of islands, that is, the southern *Sotavento* varieties, which include those of Santiago, Fogo, and Brava; the *Barlavento* varieties include those of Santo Antão, São Vicente, São Nicolau, Boa Vista, and Sal. While this grouping to some extent reflects language-external historical settlement patterns, perhaps in lieu of principled linguistic distinctions (cf. Pereira 2006), there have been observed many areal distinctions in phonological, lexical, and morphemic variation. Generally, the *Sotavento* varieties have been characterized as more basilectal, conservative, and are known to have formed at a much earlier date than the *Barlavento* varieties (Quint 2000b).

Political tensions between the interests of the *Barlovento* and *Sotavento* are manifested in the disquiet over officialization of Kriolu. Since Badiu is spoken on Santiago, the most populous

island and the seat of political power, it is the most widely spoken variety, and is typically the default variety of CVC used in media and other public domains. Many *Barlavento* speakers express concern that the Santiago variant will be imposed upon them (Coonan 2007:105-108). I argue that this issue should not preclude officialization or standardization of CVC, so long as there are genuine efforts during the officialization/standardization process to accommodate regional variation, and so long as the state is willing embrace and encourage the maintenance of local vernaculars.

Manuel Veiga (2004:114) advocates for just such a model, where a *Barlovento* standard would be based on the São Vicente variety, and a *Sotavento* standard on the Santiago variety, and that upon these standard models, dialectal variation within each respective island group could be accommodated (Coonan 2007:107-108). This may not even be wholly necessary, however, in light of the findings in Neves (2009) that suggest that the traditional north-south dialectal grouping of phonological features is not reflected in the mesolectal speech of school aged children, in whose speech the distribution of phonological features appears to reflect regional variation on an east-west axis, rather than a north-south axis.

Another ongoing point of contention is the unease surrounding Cabo Verde's historical isolation, and worries that officialization of Kriolu would cause a decline in the use of Portuguese, contribute to lower proficiency levels in the language, and thus weaken Cabo Verde's ties to other Lusophone nations, cultural traditions that have developed in Lusophone literatures and arts, and other related notions that officialization would cause Cabo Verde to become isolated from the international community (Coonan 2007:71-74). I submit that this argument is invalid on the grounds that it is a red herring fallacy. It assumes a zero-sum game by which the officialization of CVC necessarily implies a reduction in the use or proficiency of Portuguese (Coonan 2007:80-81), an outcome that is neither argued for by proponents of officialization within the public-school curriculum, nor an inevitable outcome since Portuguese is already so deeply embedded in Cabo-Verdean intuitions, not to mention the numerous examples of multilingual nation-states capable of accommodating several official languages.

Following a similar line of argument, opponents to the officialization of Kriolu point to a lack of didactic materials in the language, which they fear constitutes an insurmountable expense and logistic challenge that the state is incapable of overcoming (Coonan 2007:80-84). Advocates including Manuel Veiga (2004), Dulce Almada Duarte (1998), Marlyse Baptista (1997), and other government officials, functionaries, and intellectuals that responded to questionnaires on the topic in Coonan (2007), all advocate for a balanced bilingual curriculum in which Kriolu enjoys official status and serves as the language of instruction alongside Portuguese, and where the integration of CVC into the curriculum is guided by empirical research that would evaluate if there are tangible improvements in education outcomes (cf. Coonan 2007:74-78). Furthermore, advocates for an officialization that works towards stable bilingualism advance a compelling argument that instruction in CVC could boost comprehension and attainment in a variety of core subjects beyond language studies (Coonan 2007:77-78).

Proponents for officialization also argue that diglossia hinders Cabo-Verdeans' ability to petition their government, advocate for change, and access resources. Much of the criticism levied at language policy in Cabo Verde in Amado (2015) hinges on these issues:

> […] diglossic language policy limits the quality of democracy by way of [the] lower classes' diminished surveillatory and initiatory political participation […] diglossic language policy preserves the divide between elite and masses, whereby the latter['s] participation in politics is limited to voting (2015:vi).
>
> The linguistic divide, as such, reinforces the distance and curtails the people's participation in the processes of controlling the state. (2015:166-167).

Furthermore, the citizenry's ability to monitor and combat state corruption is hindered by the societal biases associated with diglossia in which the bureaucracy and elites have a near monopoly on access to the former colonial standard (Amado 2015:154):

> To many citizens, the state in Cape Verde is concealed because it is illegible. […] the common citizens find it quite difficult to supervise the state, its institutions and personnel. This is so because of the linguistic gap between them. […] The use of the Portuguese language, in spite of its linguistic proximity to the Cape Verdean language, the people's mother tongue, provokes a political short-circuit [...] (Amado 2015:169).

Similar impediments to the citizenry's participation in politics and ability to combat corruption, abuse, and state excess, are imposed by the Cabo-Verdean media, which on a national level is state-run (Amado 2015:178-180). Most news broadcasts are conducted almost exclusively in Portuguese, except occasionally when interviewing everyday citizens on the street. Most popularly consumed media, with exception of the widely popular Brazilian telenovelas, tends to be in CVC, particularly popular music and most interpersonal communication on social media. Cabo-Verdeans' language choice on social media platforms is a topic that is sure to be a fruitful line of inquiry for discourse-pragmatics studies, as well as studies in communications and new media.

With respect to sociolinguistic variation, there are a variety of perspectives on the status of diglossia in Cabo Verde: whether or not Kriolu is undergoing decreolization, what the extent of Portuguese influence on Kriolu has been in the modern context, the chronological order in which certain superstrate features entered the grammar or lexicon, and what superstrate-influenced restructurings might be currently underway, just to name a few.

Duarte (1998:61) and Coonan (2007:116-117) view Cabo-Verdean language policy as accelerating a process decreolization that has been underway since the end of a slavery, has been intensifying since, and that effects all domains of CVC grammar. These authors generally consider Portuguese influence on CVC to be a negative, culturally assimilative process. These authors, among others, see CVC as an essential part of Cabo-Verdean identity, they emphasize the symbolic value of CVC, particularly as it relates to resistance and anti-colonial struggle, the expressive and emotional value of the language in literature, popular music, proverbs and popular sayings (cf. Carter & Aulette 2009), and in maintaining ties with the diaspora.

Quint (2012) also describes the roles of CVC and Portuguese in Cabo Verde to be a typical example of diglossia, and that the Portuguese spoken in Cabo Verde is a non-autochthonous variety of Portuguese[24], meaning the Post-Creole Continuum model would not apply to the Cabo Verde

[24] This differs slightly from the view in Neves (2009) in her study on phonological influence from Portuguese on the speech of school-aged children. While Neves also considers the Cabo-Verdean case to be one of typical diglossia, she

case (Quint 2012). In its basilectal form, CVC is more conservative in its use of Old Portuguese and African lexical items, exhibits comparatively less morphosyntactic influence from modern Portuguese, and is spoken in rural environs. The acrolectal variety is spoken in urban contexts, and often in the diaspora, as well as among the bourgeoisie and elite classes, and exhibits greater influence from modern Portuguese (Quint 2012).

Additionally, Quint views Portuguese influence as the principle source for variation in modern CVC and shows how this variation acts on diachronic and synchronic axes. For example, he describes variation in past tense forms of the verbs *tem*/*teni* 'to have' and *bem* 'to come'. Historically these verbs were marked the regular past tense marker *-ba* that is typically applied to non-stative verbs, as in *temba*/*teneba* and *benba*, respectively. More recently, competing variants resembling the Portuguese imperfect have emerged and occur in variation with the older forms, these being *tinha* and *binha*, respectively. These variants have long been in competition in Badiu and other related varieties of CVC. But based on evidence from Izione Silva Santos (1985), it is clear that the former forms used to be much more frequent, and that in modern-day varieties they have been greatly reduced in favor of *tinha/binha* forms resembling the Portuguese imperfect (Quint 2012).

This variation, though still active, is much older than a number of newer Portuguese-derived forms that have entered urban, elite, and youth lects. Several of these verb forms resemble the Portuguese preterit, and can each be contrasted with a corresponding form that mirrors the Portuguese imperfect (9-12; from Quint 2012):

(9) *kre* 'to want'– *kis/kix/krix – kria*; Port. *quis – queria*

(10) *sabi* 'to know' – *sobi – sabia*; Port. *soube – sabia*

views the L end of the diglossic continuum as being comprised of "The Cape Verdean language", the H end of the spectrum as "European Portuguese", and the mesolect to be the "Indigenized Variety" spoken by the school-aged children in her study.

(11) *sta* ‘to be’ – *stivi*/*stevi* – *staba*; Port. *esteve* – *estava*

(12) *tem* ‘to have’ – *tivi/teve* – *tinha;* Port. *teve* – *tinha*

Most of the forms resembling the Portuguese imperfect vary with the older *-ba* marked forms, as in *kreba* and *sabeba*, while other verbs have past forms resembling the Portuguese imperfect but lack one resembling the Portuguese preterit such as *podia* and *devia,* which vary with *podeba* and *debeba*. This variation entered Kriolu more recently, perhaps with the exception of *foi,* which corresponds to the copular verb *é/ser,* which likely began to be used more widely around the same time as the imperfect forms *era, tinha* and *binha,* (Quint 2012). Notably, the use of forms resembling the Portuguese preterit only occur with ‘strong’ verbs, since ‘weak’ verbs in CVC receive a past-tense reading in their bare, simplex form. Other such examples of modern-day variation derived from European Portuguese that historically had not been integrated into CVC, and which likely occur less frequently in rural varieties, include the use of compound perfective forms with auxiliary verb *tem,* and use of the morphology associated with subjective mood (Quint 2012).

Marcia Rego (2015:95-98), on the other hand, points the difficulty and arbitrariness sometimes required to distinguish CVC from Portuguese, and vice-versa, in daily life in Cabo Verde. In popular usage, speakers of CVC do seem to distinguish a ‘continuum’ of sorts, referring to basilectal forms of CVC as ‘*Kriolu fundu’* ‘deep creole’, ‘*Kriolu levi’* ‘lite creole’ for the mesolectal or acrolectal registers, while the Portuguese itself can often be peppered with varying degrees of CVC structure or lexicon, sometimes blurring the lines between where one ‘language’ ends and the other begins, and challenging notions of whether one is witnessing a phenomenon like code-switching or code-mixing, or rather movement along a broader continuum for which the entire spectrum can be considered a single code. Ultimately, Rego (2015:66) observes, viewing CVC and Portuguese as separate codes may be futile since “Kriolu is simultaneously inside and outside of Portuguese”.

Indeed, one's view on the matter may be informed more by identity and the role of CVC and Portuguese in constructing 'Cabo-Verdeanness', rather than hard-and-fast linguistic principles. The state of diglossia in Cabo Verde, which as we have seen emerged in the late colonial period and has persisted through independence into the modern day, exerts an enormous influence on Cabo-Verdean's language choice, and which linguistic forms, expressions, or domains of knowledge are associated with Kriolu or Portuguese. For example, Rego (2015:86-89) explains that since Portuguese is the language of instruction, it is inextricably connected to the acquisition of knowledge, and to a certain extent, the possibility of social ascendancy. In a related sense it is also the language of formality, contact with the external world, and sometimes even of social aloofness (see also Coonan 2007:98-101; Carter & Aulette 2009). As Rego (2015:89-91) explains:

> [Portuguese] is spoken with 'others', that is, people from 'outside', or those of the almost equally foreign administration. People speak Portuguese with those who remain at a certain distance, hierarchically or otherwise. […] Portuguese is the language of magniloquence and effusiveness, of complicated and verbose constructions that at times seem deliberately hermetic.

CVC, on the other hand, is viewed by most Cabo-Verdeans as the language of social intimacy and directness. It is used almost exclusively in interactions among peers, family and friends. If Portuguese is the language of haute rhetoric, then Kriolu is the language of storytelling, jokes, playful banter, or deep emotional connection (Rego 2015:92-93). Bits of Kriolu can be inserted into more formal contexts as well, in brief intervals between longer stretches of Portuguese, such as during a break at work, or a meeting, or for comic relief, or for making an emotional appeal in a speech (Coonan 2007:98-101; Carter & Aulette 2009; Rego 2015:92-93). Many even suggest Kriolu is "… the best conduit of Cape Verdeans' thoughts and feelings…", that it is the medium which best conveys the 'soul' or essence of what it is to be Cabo-Verdean, and that for one to truly understand the Cabo-Verdean experience one must understand and speak it (Coonan 2007:93-94; Carter & Aulette 2009; Amado 2015:211).

CVC, as was mentioned, has a long lyrical tradition in folk and popular music, as well as poetry, refrains, popular sayings, proverbs, and other popular conceptions of Cabo-Verdean values or the emotional experience of modern Cabo-Verdeans observable in words like *sodadi* - which is associated with nostalgia and longing for one's homeland, particularly in the diaspora, - or *morabeza* – associated with the welcoming, friendly, communitarian, fun-loving character of Cabo-Verdeans (Carter & Aulette 2009). CVC is also elevated on nationalist or anti-colonialist grounds, where CVC is characterized as a language of resistance and liberation struggle (Coonan 2007:96-97). For example, during the debate over the creation of an alphabet for CVC, opponents to the implementation of ALUPEC characterized the selection of graphemes as deliberately 'anti-Portuguese'. In another example, Carter & Aulette (2009) find in their ethnographic study of CVC as viewed in popular Cabo-Verdean culture, that women employ CVC proverbs

> […] as tools for exposing and challenging gender inequality and gender injustice. Women draw on proverbs and sayings to illustrate what they are not getting out of a relationship, work situation or the society as a whole. Women employ proverbs and sayings as instruments which reflect their needs, concerns, ideas and feelings.

In this subsection, I explored the role of CVC and Portuguese play in modern Cabo Verde. I considered how the social role associated with each language has been influenced by historical trends, and how in-turn the social prestige associated with each language informs social relationships, socioeconomic status, and Cabo-Verdeans means for constructing and projecting national, regional, or class identity. The implications of these social factors on linguistic variation were also considered, in particular the chronology of when superstrate features entered CVC grammar, and the nature of diglossia or the relevance of the 'Post-creole continuum' model for explaining the sociolinguistic dynamic in Cabo Verde.

2.7 CHAPTER CONCLUDING REMARKS.

In this chapter, I began with the historical setting in which Cabo Verde was settled and the first 150-200 years of the development of society on the island of Santiago. I concluded that the

demographic and sociohistorical development of society on Santiago was the earliest prototype for the slave-based settlements founded throughout the Atlantic basin over the course of the colonial era, from which so many restructured contact languages emerged. Santiago also became an early central node in the trans-Atlantic Slave Trade.

With respect to linguistic developments, I reviewed the 'origins' debate for the UGCP group of languages, concluding – following Parkvall (2000), Jacobs (2010, 2012), and Jacobs & Quint (2016) – that the preponderance of the historical, linguistic, and population genetic evidence, conspires in favor of a Santiago-birth hypothesis. The Santiago-birth hypothesis states that a proto-UGCP vernacular emerged on Santiago and was later dispersed to the UGC mainland and the rest of archipelago, where each variety embarked on its own developmental trajectory, resulting in the contemporary varieties of the UGCP (Jacobs 2010).

I then returned to the extra-linguistic history of Cabo Verde, from the stagnation and suffering during the remainder of the colonial era, to the liberation struggles of the 20th century, and through independence and the development of a multiparty parliamentary republic. Finally, I addressed the sociolinguistic status of CVC, particularly with respect to language ideologies and identity, and how these inform various questions such as orthography, standardization, officialization, and morphosyntactic variation.

Having established a historical, sociological, and diachronic/typological base upon which develop our understanding of sociolinguistic variation in CVC SPE, in the next chapter, I turn to a review of a historical reconstruction of the proto-CVC subject pronoun inventory conducted in Lang (2012). This is followed by a brief summary of the subject domain and subject-domain-adjacent properties of CVC's primary source languages: Late Medieval/Early Classical Portuguese, Wolof, and Mandinka.

Chapter 3: An overview of historical reconstructions of the Cabo-Verdean Creole subject pronoun system and the subject domain in its major source languages.

The objectives of this chapter are to provide a sketch of the historical formation of Cabo-Verdean Creole (CVC henceforth) subject pronoun system, as well as an overview of the subject-domain (and some subject-domain-adjacent morphosyntactic characteristics) of its primary source languages: Late Medieval/Early Classical Portuguese, Wolof, and Mandinka. With respect to this first objective, I will give a synopsis of Lang (2012) in which the author conducted a diachronic reconstruction of the CVC subject pronominal inventory. This reconstruction is a detailing of the transformational steps in the diachronic development of the CVC subject pronoun inventory, but does not differ greatly in its assessment of the origin of CVC subject pronouns from the analysis offered in Quint (2000a) (although the authors differ in their assessment of substrate influence in the system). In continuation of the convention that I adopted in the previous chapter I refer to this early stage in the development of CVC as 'proto-CVC'[25].

Next, I conduct a review of the literature on the major features of the subject domain and related aspects of morphosyntax in Late Medieval/Early Classical Portuguese, the language that was the superstrate on Santiago during the first 100 years following the settlement. It is noted that Late Medieval/Early Classical Portuguese was a consistent Null Subject Language and had a verb-second restriction that limited verb placement to the second position in the clause (but not necessarily in linear position, see discussion below).

I then briefly discuss the overt subject pronoun systems of CVC's two primary substrate languages, Wolof and Mandinka. This is a discussion of the modern subject pronominal inventories of these languages, since, regrettably, there is little to no historical record or any reconstructions of the properties of these languages as they were spoken in the 15th and 16th centuries. As was discussed in the prior chapter, it is in the 15th and 16th centuries that speakers of

[25] Lang uses the term *santiaguense primitivo* 'Primitive Santiago [Creole]'.

Wolof and Mandinka were at their greatest numbers on Santiago, and thus when the features of these languages were active in the 'feature pool' for proto-CVC. Despite the lack of records for the varieties of Wolof and Mandinka that were contemporaneous with proto-CVC during its formative early stages, I suggest that even analyses of the modern varieties of these languages will reveal a striking similarity with modern CVC.

3.1 PRIOR RECONSTRUCTIONS OF THE CABO-VERDEAN CREOLE SUBJECT PRONOUN SYSTEM .

Quint (2000a) and Lang (2012) take on the task of reconstructing the historical subject pronoun inventory of *Badiu* (CVC from Santiago) and *Foguense* (CVC from Fogo), the two oldest varieties of CVC. Lang relied on available grammars of the superstrate and substrates languages to conduct his diachronic reconstruction, since there is no metalinguistic documentation[26] of CVC dating before 1880[27] (2012:20). Since the present study analyses data primarily from the Santiago variety of CVC, I focus here only on his reconstruction of the subject pronouns for Badiu. Before any discussion of the proto-CVC subject pronoun inventory, it will be useful for referential purposes to recast the contemporary CVC subject pronoun inventory; Lang (2012:22) adopts a bipartite tonic-atonic classification and assumes that disyllabic *a-* initial forms are reserved for use these as topics (Table 12).

[26] Prior written attestations of CVC come in the form of observations about its existence rather than any metalinguistic documentation or analysis.

[27] The first known written grammar of CVC was published in 1887 by António de Paula Brito (see Quint 2008c) and contains the earliest account of the subject pronoun system available today.

Table 12. The subject pronouns of contemporary Badiu (CVC from Santiago) (adapted from Lang 2012:22)

Person-number	Tonic	Atonic
1SG	*(a)mi*	*N*
2SG	*(a)bo*	*bu*
3SG	*(a)el*	*e(l)*
1PL	*(a)nos*	*nu*
2PL	*(a)nhos*	-[28]
3PL	*(a)es*	*es*

The primary historical superstrate for proto-CVC was Late Middle/Early Classical Portuguese. I adopt the term Late Medieval/Early Classical Portuguese following Lang's (2012:22-23, and references therein) terminology, *"português tardo-medieval e clássico"*. The period between the 15th-16th centuries was a transitional period in the history of the Portuguese language. Usually, the stage of the language between the 13th and 16th centuries is known as Old Portuguese, but this period can itself be subdivided into the Galician-Portuguese stage from the 12th to the mid-14th century, followed by the *Middle* or *Late Medieval Portuguese* stage (cf. Castro 2006; Galves & Paixão de Sousa 2017:153$_{n2}$). Galves & Paixão de Sousa (2017:152) adopt the term "Classical Portuguese" for the phase of the language between the 16th and 18th centuries.

The Late Medieval/Early Classical Portuguese subject pronoun inventory was well documented. Lang's (2012:22-23) representation was drawn from historical grammars such as those of Teyssier (1980), Maia (1986), Riiho (1994), and Silva (2008) (Table 13).

[28] Lang claims that "Na segunda pessoa do plural, não há formas átonas: usa-se em todas as posições a forma tónica *nhos* ['ɲos]." [In the second person plural, there are no atonic forms: the tonic form *nhos* ['ɲos] is used in all positions.]. However, as we saw from the table adapted from Baptsita, Mello, & Suzuki (2007) (Table 4, Section 2.3.1; see also Table 19, Chapter 4.1), and as we will see from other accounts, some researchers do assume an atonic *nhos/nhas*.

Table 13. Late Medieval/Early Classical Portuguese [tonic] subject pronouns (adapted from Lang 2012:22-23)

Person-number	Subject pronoun
1SG	*(eu) mim*
2SG	*vós*
3SG m.	*(ele), el*
3SG f.	*éla*
1PL	*nós*
2PL	*vós*
3PL m.	*(eles), eis*
3PL f.	*élas*

The 1st person singular subject *mim* is noted in Maia (1986:663), who claims that it "varied sporadically" with *eu*. Note the alternative form *el* for the 3rd person singular (possibly by apocope or reduction of the final vowel) and *eis* for 3rd person plural. 2nd person singular *vós* was the standard form of address in relationships of social distance, as opposed to *tu*, which was employed in cases of a more intimate relationship between speaker and interlocutor (Cintra 1986). For this reason, Lang (2012:24-25) assumes that *vós* would have been the standard form of address directed at the enslaved inhabitants of Santiago. While *tu* certainly existed as a form of address in Late Medieval/Early Classical Portuguese, it left no morphological trace in contemporary Badiu (Quint 2000a:180-182). The form of address *vossa mercê*, which eventually developed into modern *você*, did not come into usage until the last half of the 15th century, and was at first restricted to use with the heads of the royal family. *Vossa mercê/você* did not spread as a form of address directed at other members of society until later and thus would not have been present in Santiago during the formative stages of proto-CVC.

On the basis of the Late Medieval/Early Classical Portuguese pronoun inventory, Lang (2012:26) posits a reconstruction of the proto-CVC subject pronoun inventory (Table 15).

Table 15. Early Proto-CVC subject pronouns (adapted from Lang 2012:26)

person-number	tonic	atonic
1SG	*mim* > *mim* ['mĩ]	*mim* > **mim* ['mĩ]
2SG	*vós* > **bós*	*vós* > **bos*
3SG	*el(e)* > *el*	*el(e)* > *el*
1PL	*nós* > *nós*	*nós* > **nos*
2PL	*vós* > **bós*	*vós* > **bos*
3PL	*eis* > **eis*	*eis* > *es*

The pronoun used by the Portuguese speaking colonists > the early Badiu pronoun, reconstructed [unattested] forms are marked with an asterisk.

Lang posits that proto-CVC tonic pronouns came about from regular sound changes applied to Late Medieval/Early Classical Portuguese tonic subject pronouns. Atonic proto-CVC subject clitics (SCs henceforth), as we will see, follow a less-straightforward developmental path, but at least some of them can be tied directly to regular sound changes as well. For example**,** CVC prohibits open vowels [ɛ], [a] or [ɔ] in atonic syllables. For the SCs that were directly derived from preexisting tonic forms (3rd person forms, see discussion ahead) the open [ɛ] of the tonic subject pronouns were closed to [e] in SCs[29]. With respect to early proto-CVC 2nd person SC *bos*, the labiodental changed to [b], another regular sound change that affects core lexical items in CVC. 3rd person *eis* experienced monopthongization, yet another widespread sound change in CVC affecting Late Medieval/Early Classical and Modern Portuguese */ei/* diphthongs.

Quint (2000a:262) in his own reconstruction of a proto-variety of Upper Guinea Creole Portuguese - which he takes to be the predecessor or source-variety for Papiamentu, Guinea-Bissau Creole, and CVC - posits *mi, bo, el, nos, bos* as the available pronouns in this proto-variety. He

[29] This is indicated by the removal of the diacritic [´] from the vowels in the "atonic" column in Table 13.

further proposes that *a-* initial forms existed in emphatic constructions, but given the lack of [atonic] SCs in contemporary Papiamentu, he concludes that proto-Upper Guinea Creole Portuguese likely lacked unstressed subject pronouns as well. Presumably, atonic forms could be derived from a process of grammaticalization on these tonic forms in the proto-vernacular (a proposal to which I return in future chapters). While this distinction could potentially have implications with respect to Lang's claims regarding substratal influence, it is unclear if Quint and Lang are referring to the same early developmental stage of proto-CVC.

The next phase in the development of proto-CVC subject pronouns comes from the first available documentation of the system in the last quarter of the 19th century in the grammar by António de Paula Brito. In this subject pronoun inventory (Table 16), white cells contain pronouns that align with the expected developmental outcome given the early system postulated in Table 13, light grey cells contain pronouns that represent intermediary stages or changes in progress towards their corresponding modern CVC pronouns, and dark grey cells indicate unexpected or innovative changes.

Table 16. Late 19th-century CVC subject pronouns (adapted from Lang 2012:27; itself from Brito 1967[1887])

Person-Number	Tonic	Atonic
1SG	*(a)mĩ*	*ĩ*
2SG	*(a)bó*	*bu*
3SG	*(a)él*	*ê*
1PL	*(a)nós*	*nu*
2PL	*(a)nhôs*	-
3PL	*(a) ēss*	*ēss*

The initial *a-* disyllabic forms have now become an active variant. 3rd person plural *ēss* indicates a lengthening of both the vowel and consonant, presumably an intermediate stage in the changes *eis* > *ēss* > *es.* Late Medieval/Early Classical Portuguese second person plural *bós* came to be replaced by *(a)nhos* (PL) and singular tonic-atonic pairs *(a)nho* - *nhu* and *(a)nha* – *nha* (not

listed here), all of which have their origin in *senhor(es)*, an observation most evident in the archaic and rural variants *nhor* (M), *nhara* (F), and *nhoris* (PL) (Lang 2012:30-31).

1st person plural *nu* from *nós* may seem an odd outcome given that CVC makes extensive use of plural morpheme *-s*, and coda /s/ elision is not a regular phonological process typically associated with this language. Lang (2012:28) considers *nu* to be a straight-forward transference of the nearly homophonous Wolof 1st person plural atonic person marker *nu*. The interpretation of coda /s/ in Late Medieval/Early Classical Portuguese tonic *bós* as a plural marker could have led to coda /s/ being dropped giving 2nd person *(a)bó*, whereas unstressed *bu* could be the result of analogy with 1st person plural tonic-atonic pairing *(a)nos* - *nu* (Lang 2012:28,30).

Lang (1999, 2012:27,27$_{n7}$) observes that while Brito chose to represent the unstressed 1st person pronoun as *ĩ* in his table, his descriptions of the use and actual phonetic realization of the pronoun indicate that its associated phoneme is likely similar in status to the modern Badiu 1st person subject clitic (orthographically represented in the ALUPEC alphabet as *N*) in that

> o significante deste pronome não é um fonema ou uma sequência de fonemas, mas o simples traço da nasalidade. Traço que se combina com a consonante inicial da palavra seguinte, tornándo-a pré-nasalizada, e com a vogal da palavra precedente, tornándo-a nasal. [the signifier of this pronoun is not a phoneme or a sequence of phonemes, but [rather] a simple trace of nasality. [This] trace combines with the initial consonant of the following word, making it pre-nasalized, and with the [final] vowel of the prior word, making it nasal.].

There are three possible ways that *N* might have come about: (i) it is possible the nasal trace could be derived from a series of reductions to unstressed Late Medieval/Early Classical Portuguese *mim* [*mĩ*]; (ii) this might constitute a case of morphological convergence between *mim* and the 1st person singular subject pronoun in the Manding languages, which in Bambara (and Mandinka, see Section 3.4 below) is orthographically represented as <*ń*>, and which is phonetically and phonologically similar in value to the description given for the CVC SC, but with the Bambara diacritic representing a high tone; (iii) there was no convergence, instead *N* is the result of direct borrowing or imposition from the substrate. This final viewpoint is the one adopted by Lang (2012:29).

One last process of analogy would finally yield the contemporary CVC subject pronoun inventory (Table 17). Notice that in Table 15 the tonic 2nd person singular *(a)bó* and 1st person plural *(a)nós* have diacritics indicating open vowel [ɔ]. These vowels would eventually become closed [o], purportedly via analogy with *(a)nhos* [ɐɲos] and *(a)nho* [ɐɲo] (Lang 2012:31). Though Lang does not list them, Quint (2000b:164) observed unstressed *nhu* (M) and *nha* (F), which complete the analogical regularizations associated with the *(a)nos – nu* and *(a)bo – bu* pairings.

Table 17. Contemporary Badiu subject pronouns.

	Tonic	Atonic
1SG	*(a)mi*	*N*
2SG	*(a)bo*	*bu*
3SG	*(a)él*	*e(l)*
1PL	*(a)nos*	*nu*
2PL	*(a)nhos*	-
3PL	*(a)es*	*es*

Light grey cells indicate completion of changes in progress from 19th century, dark grey cells indicate innovative change by analogy.

Thus far, the development of the CVC subject pronoun system was explored, and it was concluded based on Quint (2000a) and Lang (2012) that this development involved a combination of regular sound changes, analogical changes, and likely contact-driven substrate-superstrate cross-linguistic convergence (Quint p.c.). Lang (2012:32), who adopts a stronger, substrate-imposition analysis for the origin of the first-person atonic forms, considers these developments to be "*claramente un ato de crioulização*" [clearly an act of creolization]. Lang's usage of the term 'creolization' seems to also reflect the notion that the morphophonological shape of most of CVC's subject pronouns is based mostly on Late Medieval/Early Classical Portuguese subject pronouns, but the underlying structure, with its bipartite tonic-atonic opposition, reflects the makeup of the Mandinka and Wolof systems, which are also disjunctive on the basis of stress (Lang 2012:32).

Regarding this latter observation, it is possible that this too could be taken to result from contact-driven cross-linguistic convergence, since we know that the oblique pronominal system of CVC's was also disjunctive on the basis of stress. Further, it is well known that the development of bipartite stress-disjunctive nominative pronominal inventories was a diachronic outcome for many Gallo-Romance, Gallo-Italic, and Rhaeto-Romance vernaculars (suggesting that such structural developments can come about without necessitating a substrate language bearing the equivalent organizational properties; bipartite subject pronoun inventories in Romance are discussed further in Chapter 4).

This subsection explored a historical reconstruction of the proto-CVC subject pronoun inventory and its development into the subject pronouns of contemporary CVC, as was carried out in Lang (2012). This system eventually came to include two morphemes that appear to have come about from cross-linguistic convergence between superstrate *nós* and *mim* with Wolof origin *nu* for the atonic 1st person plural SC *nu*, and Mandinka/Bambara-origin *N* for the 1st person singular SC, respectively. All the other modifications that obtained in the development of CVC subject pronouns appear to have proceeded from the internal processes related to vowel quality, length, and analogical change (Lang 2012:33-34). Such changes might be viewed as resulting from a broader process of grammaticalization on tonic forms (an assertion that is revisited in Chapter 7). I turn now to a brief review of the literature on Subject Pronoun Expression in CVC's source languages, starting with Late Medieval/Early Classical Portuguese.

3.2 Subject Pronoun Expression and related phenomena in Late Medieval/Early Classical Portuguese.

In this subsection, I will briefly review the primary characteristics of the subject domain and subject-domain-adjacent morphosyntactic properties of Late Medieval/Early Classical Portuguese. The objective is to establish what morphosyntactic features would have been present in the superstrate language to CVC during the early formative stages of proto-CVC. This will

contextualize the degree of morphosyntactic restructuring that took place as proto-CVC began to take shape. It will also allow for possible identification any diachronic processes that may have contributed to the dynamics of Subject Pronoun Expression in modern-day CVC.

Before addressing the features of the subject domain and adjacent domains in Late Medieval/Classical Portuguese, it would be worthwhile to reiterate the historical demarcation between Old Portuguese and Late Medieval/Classical Portuguese as it relates to the early formative period of proto-CVC. Recall that in Chapter 2, it was established that proto-CVC is likely to have consolidated rapidly following the settlement of Santiago (1460s). A vernacular discernably distinct from the Late Medieval/Early Classical Portuguese spoken by the 15th century European settlers had likely become mostly consolidated sometime before the last quarter of the 15th century (Quint p.c.).

Santiago remained an important node in the trans-Atlantic Trade and continued to grow until the early 1600s, when the island entered a long period stagnation, decline, isolation, and turmoil that lasted at least until the climax of *companhia* era in the 18th century. During this period of stagnation following the decline in the first quarter for the 17th century, input from contemporaneous varieties of Portuguese would have been quite limited, since the number Portuguese settling in or traveling through Cabo Verde would have been low (Patterson 1988).

As was briefly noted, the traditional division of Portuguese into historical periods (if perhaps somewhat arbitrary or contingent on language-external historical circumstance) situates the beginning of Old Portuguese in the 12th century, lasting until the 15th or mid-16th century. The first half of this period is usually associated with Galician-Portuguese ending around the mid-14th century. The second period is associated with Middle Portuguese or Late Medieval Portuguese. After Late Medieval Portuguese came Classical Portuguese, which itself can be considered a transitional period between Old Portuguese and modern European Portuguese (Teyssier 1980, 1997; Galves & Paixão de Sousa 2017; Spina ed. 2017).

The early formative period of proto-CVC, then, coincides with the final years of Late Medieval Portuguese as it was transitioning into Early Classical Portuguese. The description in

this subsection will address the findings from a research program established mostly on the basis of analyses of a corpus consisting of texts from the Classical Portuguese period. Throughout the remainder of this subsection, I urge the reader to keep in mind that the data presented reflect a historical period for which only the first century coincided with the formative period of proto-CVC on Santiago, and reflects a literary register rather than the oral vernacular used on Santiago in the 15th and 16th centuries. That said, Old Portuguese and Early Classical Portuguese shared many morphosyntactic properties related to verb and subject placement (Ribeiro 1995; Galves & Paixão de Sousa 2017), though there were notable differences in clitic placement (Martins 1994; Namiuti 2008). Given the paucity of data regarding the superstrate Lusophone vernacular present on Santiago in the 15th and 16th century, the analyses of Classical Portuguese Subject Pronoun Expression described in this subsection will have to serve as a best possible approximation. Given the transitional nature of the superstrate during this period, and for the sake of consistency and clarity, I will continue to use the term Late Medieval/Early Classical Portuguese throughout the remainder of this chapter.

The defining characteristic of the Late Medieval/Early Classical Portuguese subject domain that distinguishes it from modern European Portuguese was its structural affinity to verb-second languages (V2 henceforth). V2 properties have also been attested in other varieties of Old Romance (Benincà 1984; Ribeiro 1995; Salvi 2000; Sitardou 2012), including in various studies on the loss of V2 properties from Old to Middle French (Adams 1987; Roberts 1993; Vance 1997). Like Old Romance, Late Medieval/Early Classical Portuguese was also a *pro*-drop language or consistent Null Subject Language, meaning that in addition to V2 features it also allowed for abundant null subjects (Galves & Paixão de Sousa 2017:155). Much of the contemporary line of research on V2 in Late Medieval/Early Classical Portuguese was initiated by investigations into object clitic placement (cf. Martins 1994; Torres Moraes 1995; Galves 1996) and was championed in previous decades by several researchers who made ample use of the *Tycho Brahe Corpus*. This syntactically annotated corpus consists of historical Portuguese texts written by authors born

between 1502 and 1836 (Paixão de Sousa 2004; Galves, Britto, & Paixão de Sousa 2005; Galves & Paixão de Sousa 2017, *inter alia*).

In the formal syntactic literature, V2 has long been considered a result of movement of the verb to the C (node of the complementizer phrase layer) (Den Besten 1983; Antonelli 2011:503; Andrade 2015; Galves & Paixão de Sousa 2017:153). Another feature of V2 languages is that there is no pre-verbal position dedicated solely to the subject, though a wide variety of constituents, including subjects, can be appear in a [pre-verbal] topic position given the right pragmatic-discursive context (Torres Moraes 1995; Galves 1996; Paixão de Sousa 2004; Antonelli 2011; Andrade 2015; Galves & Paixão de Sousa 2017:153). As can be observed in (13), the pre-verbal constituent can be a subject (13a), some verbal argument other than the subject (13b), or an adjunct (13c) (adapted from Antonelli 2011:504; itself from Ribeiro 1995).

(13) a.) *O honrado padre São Beento* ***deu*** *todalas cousas.*
DET honorable father saint B. give.PAST.3S all-DET things
'The honorable Father Saint Benedict gave all of the things.' (Ribeiro 1995:113)

b.) *E desto se* ***nembrou*** *el*
and from-this CL.3S.REFL name-PAST.3S 3S.MASC
'And he named himself after this' (Ribeiro 1995:114)

c.) *Com tanta paceença* ***sofria*** *ela esta enfermidade*
with such patience suffer.IMP.3S 3S.FEM this illness
'She suffered this illness with such patience' (Ribeiro 1995:114)

Additionally, in Late Medieval/Early Classical Portuguese and other V2 languages, fronting of objects to a pre-verbal position triggers OVS order (visible in surface form when the subject is overt). In Antonelli's (2011:506-507) analysis of data from the Tycho Brahe Corpus,

objects dislocated to the left periphery of the phrase categorically triggered a post-verbal subject (14a-e).

(14) a.) *Notavel informação* ***deu*** *este Espirito em poucas palavras.*

notable information give.PAST.3S this spirit in few words

'This Spirit gave notable information in few words.' - Maria do Ceu

b.) *Tudo* ***entendia*** *a Sobrinha*

everything understand.IMP.3S DET niece

'The niece understood everything.' - Maria do Ceu

c.) *poucas saudades* ***teria*** *este predistinado espirito de taes*

little longing have.COND.3S this predestined spirit of such

grandezas

greatness

'This predestined spirit of such greatness would have had few longings.'

–Maria do Ceu

d.) *Tudo isto* ***continha*** *o papel daquele nobre Senado*

all of this contain.IMP.3S DET paper of-that noble senate

'The paper of that noble Senate contain all of this.' – André de Barros

e.) *Muito, do que se tem dito das gentes, que*

much of-DET COMP CL.IMPRS AUX.3S say.PCP of-DET peoples COMP

por aquele Sertão demoram, e bebem em tão dilatado rio,

around that interior loiter.3P CONJ drink.3P in such distended river

avaliam *algumas Histórias por fabuloso.*

assess.3P some stories as mythical

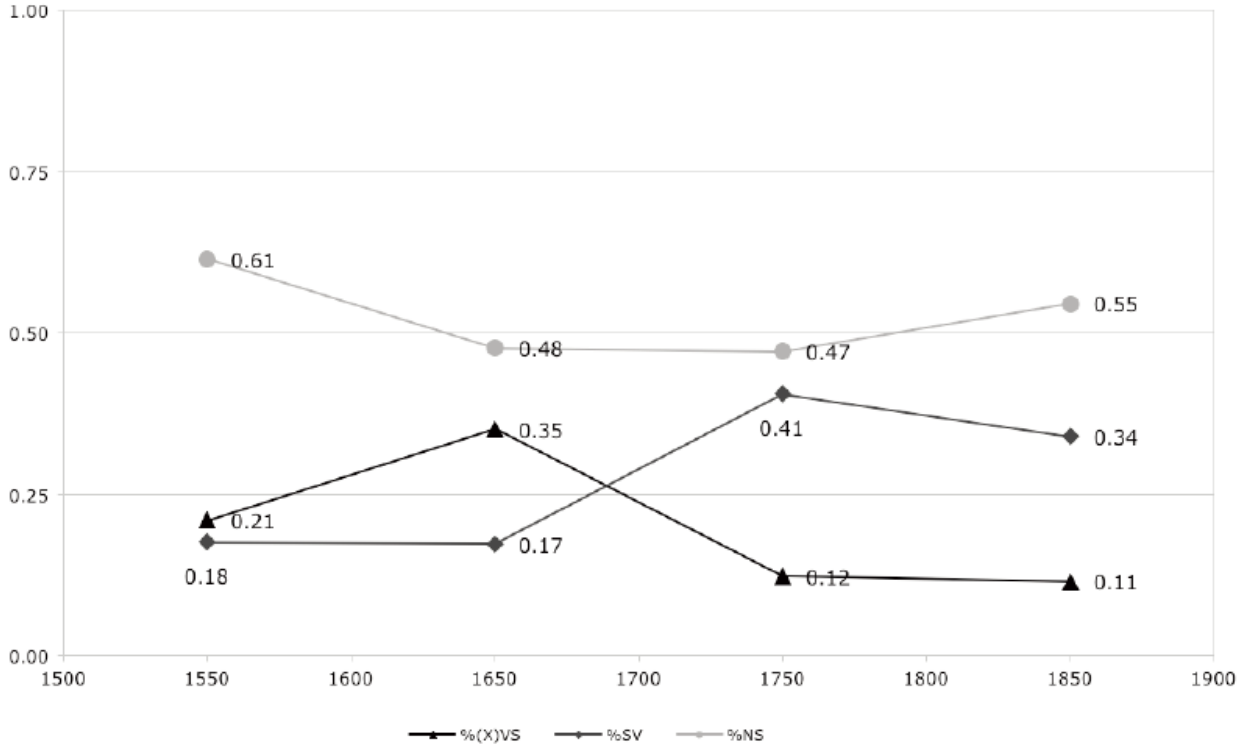
'Some stories assess as mythical much of what has been said about the peoples who loiter around those backcountries and who drink from such a distended river.' – André de Barros.

Further evidence for V2 in Late Medieval/Early Classical Portuguese came from examining the placement of object clitic pronouns. V1 clauses categorically trigger enclisis, while in nondependent V2 clauses there is variation between proclitic and enclitic placement (Martins 1994; Torres Moraes 1995; Galves 1996; Paixão de Sousa 2004; Galves *et al.* 2005; Galves & Paixão de Sousa 2017:153).

In one of the recent studies that made use of the Tycho Brahe Corpus, Galves & Paixão de Sousa (2017) compiled 34,293 tokens to examine the interactions between the V2 feature and clitic placement, prosody, and Subject Pronoun Expression. One of the most immediately striking differences between Late Medieval/Early Classical Portuguese and the contemporary European Portuguese subject domains is in surface word order. Glaves & Paixão de Sousa (2017:157) found that, from the 16th to the 19th century, rates of (X)VS order (excluding null subjects, which the authors assume to be post-verbal) occurred at higher rates than SV in the 16th century and 17th centuries.

In the 18th century, these rates flipped, and SV surpassed (X)VS as the most frequent order of realization of an overt subject. Null subject realizations vary by a range of 12% throughout this period, but remain comparatively stable, and consistently occurred at higher rates than either possible positional realization of an overt subject (Figure 4). It would seem, then, that *pro*-drop status remained constant during the transition from Late Medieval/Early Classical Portuguese to modern European Portuguese, and likely had been stable for long before that. This is irrespective of shifts in patterns of overt subject placement and word order (Glaves & Paixão de Sousa 2017: 157).

Figure 2. “The expression of subjects in main clauses, sixteenth to nineteenth centuries, by 100-year periods” (Glaves & Paixão de Sousa 2017:157)



“Proportions of NS: null subjects, SV: preverbal subjects, (X)VS: post-verbal subjects”

Given these data on transformations in Late Medieval/Early Classical Portuguese, we can assume that the subject domain in the superstrate to proto-CVC was a V2 + *pro*-drop system. Major changes in this superstrate system appear to have come well after the boom-and-bust demographic and economic cycles had run their course on Santiago; by that time, CVC would long since have consolidated, and input from contemporary varieties of European Portuguese would have been minimal (again, because there were so few Portuguese transiting through or settling in Cabo Verde during that time; cf. Patterson 1988). However, it should be considered that changes in the vernacular may have come about more quickly than what is reflected in texts compiled in the Tycho Brahe Corpus.

To summarize, under the V2 + *pro*-drop system, movement of V to C is the primary formal mechanism that generates surface word order in which the verb can appear in V1, V2, or V3 linear

positions, given that the left periphery makes available numerous slots for topicalized or focalized pre-verbal constituents, which can include various elements in addition to subjects. The post-verbal position can host focalized or non-focalized subjects and constitutes the canonical position for the early Late Medieval/Early Classical Portuguese lexical subject. Finally, major shifts in overt subject expression and word order had little effect on the *pro*-drop status of Late Medieval/Early Classical Portuguese, which, with its stable rates of null subject as the most frequent subject type, should properly be classified as a consistent Null Subject Language in which inflectional agreement morphology is responsible for the identification of *pro.*

Like in contemporary European Portuguese, V2 features were lost during the restructuring of proto-CVC, and compared to European Portuguese, post-verbal subject became even more limited; post-verbal subjects occur only with certain unaccusative verbs and in instances of focalization in contemporary CVC. Unlike Late Medieval/Early Classical Portuguese and contemporary European Portuguese, as we will see ahead, rates of null subject have been greatly reduced in CVC, and the contexts required for their identification have become more restricted. This is presumably because of the infinitival origin the CVC verb (cf. Quint 2008b)[30].

In the next subsection, I turn to brief review of the subject domain in the two primary substrate languages to proto-CVC, Wolof and Mandinka. Doing so will allow for further comprehension of the extent of morphosyntactic restructuring during the early formative stages of CVC and for the identification of any possible substrate morphosyntactic features that might have been retained in the subject domain of contemporary CVC.

3.3 THE PRIMARY SUBSTRATES.

Given the lack of documentation of the 15th and 16th century vernacular varieties of Wolof and Mandinka - the substrate languages to proto-CVC during its early formative period - any consideration of the subject domain in these languages will have to be based on contemporary

[30] An alternative proposal could be the loss of bound person-number subject-verb agreement morphology during the early restructuring process.

accounts (see Quint 2000b:24-26 for more on this issue). As was noted in Chapter 2, like CVC, Wolof and Mandinka both feature a disjunctive tonic-atonic opposition in their subject pronominal inventories, an observation that is indicative of the extent of substratal influence, or cross-linguistic convergence, observable in the proto-CVC subject domain.

3.3.1 Some aspects of the subject domain in Wolof.

In Wolof, subject pronominals share a morphological paradigm with other functional nominals that act as locative markers, as well as object pronominals, and these overtly mark person and number. The entire paradigm of functional nominals can be divided into strong stressed forms and weak unstressed forms (Table 18) (Zribi-Hertz & Diagne 2002). In this subsection I will focus strictly on the atonic subject pronominals or 'subject markers', which are roughly analogous to what we have thus far been calling subject clitics[31]. In the literature on Wolof, atonic subject pronominals are generally referred to as 'subject markers' or 'person markers' due to some of their inflectional properties (and theoretical-paradigm-specific conventions) (Dunigan 1994), though as will see ahead this assessment is contested.

[31] As has been mentioned and will be discussed in more detail in the coming chapters, the morphosyntactic category and underlying structural position claimed for SCs/person markers is contested in several languages, including Wolof.

Table 18. “Wolof person and locative markers” (adapted from Zribi-Hertz & Diagne 2002:844)

Person-number/Function	Tonic	Atonic
1SG	*man*	*(m)a*
2SG	*yow*	*nga,ya*
3SG	*moom*	*Ø, (m)u*
1PL	*nun*	*ñ(u)*
2PL	*yeen*	*Ngeen, yeen*
3PL	*ñoom*	*ñ(u)*
LOC, -dx, -obv	*cii*	
LOC, -dx, +obv	*caa*	
LOC, +dx, -obv	*fii*	
LOC, +dx, +obv	*faa, foofu*	

“functional locatives are specified for obviation (obv) and deixis (dx)” (Zribi-Hertz & Diagne 2002:844)

To begin with, it should be noted that all strong stressed forms are bimorphemic (and all but 2nd person singular *yaw* are disyllabic), while all weak unstressed forms are monomorphemic, and most of the strong forms are based on or have some morphophonological resemblance to their weak unstressed counterpart (Torrence 2005:33). Several dialects of Wolof realize the first person plural strong and weak pronouns with an alveolar nasal rather than a palatal, yielding *noom* and *nu*, the latter of these being the proposed origin of the nearly homophonous CVC 1st person plural SC *nu* suggested by Lang (2012).

The distribution of strong stressed pronominals resembles those of CVC and other languages with strong pronominals in that they can occur in the same position as any lexical DP, as the object of a preposition (15a), a left dislocation (15b), or as an object focus constituent (15c) (Zribi-Hertz & Diagne 2002:844). Notice that (15b) resembles the amalgamated double-subject

construction in CVC, a structure that we will see is available in several languages with atonic SCs/person markers.

(15) a.) *Moodu liggey ø ø ø -na ø ak* ***ñoom***
Moodu work -IMPF -PAST -NEG +F 3.SG with 3PL.STR
'Moodu has worked with them'

b.) ***ñoom*** *ñu -a (>ñoo) lekk ø ø ø ceeb -bi*
3PL.STR 3PL.CL COP eat -IMPF -PAST -NEG rice DET.DEF.SG
lit. 'They/them, it is they who at the rice'

c.) ***ñoom*** *-la Moodu gis ø ø ø*
3PL.STR COP Moodu see -IMPF -PAST -NEG

Wolof subject markers experience complex variations in both surface form and linear position depending on the clause and on the morphophonological context (Torrence 2005:25-28). They can undergo right attachment, left attachment, or remain unattached, though even in cases of right attachment they are never 'true' proclitics, since Wolof is an initial-stress enclitic-only language (Zribi-Hertz & Diagne 2002:847-849). Given the stress-initial pattern, a Wolof subject marker in initial position can receive default stress, but in any other position it remains unstressed and phonologically attaches to weak morphemes.

Thus far in the discussion of Wolof nominative elements, I have been alternating between referring to the atonic elements in Table 16 as 'subject clitics' and 'subject markers'. This reflects a disputed classification in the literature as to the underlying clausal position and morphosyntactic status of these elements. While accounts such as Dunigan (1994) understand Wolof atonic nominative pronominal elements to be true 'syntactic clitics' that arrive in their surface position via Head Movement, more recent accounts take these to be 'subject markers' or 'person-inflection markers' (cf. Zribi-Hertz & Diagne 2002; Torrence 2005; Dione 2013). Researchers adopting the latter perspective suggest that Wolof subject markers can play dual roles, one as subject person

inflection markers in V-to-INFL constructions (~tense heads/verbal affixes) (16a, 16b, and 16c), or, as 'true' argumental nominative pronominals (~phonological subject clitic/unbound anaphoric pronoun) (17a-b).

(16) a.) V-to-INFL

xale -yi lekk ø ø ø -na ***ñu*** *ceeb -bi*

child DET.DEF.PL eat -IMPF -PAST -NEG +F 3PL.CL rice DET.DEF.SG

'The children have eaten the rice.'

b.) Imperfective-to-INFL

xale -yi di ø ø -na ***ñu*** *lekk ceeb -bi*

child DET.DEF.PL +IMPF -PAST -NEG +F 3PL.CL eat rice DET.DEF.SG

'The children will eat the rice.'

c.) Explicative-to-INFL

xale -yi da ***ñu*** *lekk ø ø ø ceeb -bi*

child DET.DEF.PL EXPLIC 3PL.CL eat -IMPF -PAST -NEG rice DET.DEF.S

'(It is because) the children have eaten the rice.'

(17) a.) Object focus construction

xale -yi *ceeb -bi -la* ***-ñu*** *lekk*

child DET.DEF.PL rice DET.DEF.SG COP 3PL.CL eat

'The children, it is the rice, which they ate.'

b.) Subject focus construction

xale -yi, ***ñu*** *-a lekk ceeb -bi*

child DET.DEF.PL 3PL.CL COP eat rice DET.DEF.S

'The children, it is they who ate the rice.'

The constructions in (16a-c), are identified by Zribi-Hertz & Diagne (2002:857) as "saturated finite clauses, i.e., independent clauses specified for the full range of inflectional

features […]: Finiteness, Aspect, Tense, and Polarity." Zribi-Hertz & Diagne specify further that these are the only constructions in Wolof where the "subject marker occupies a linear position which is closed to a lexical subject […] the lexical subject (*xale-yi*) is construed as neutral (i.e., untopicalized, unfocalized)." On the other hand, in (17) the subject marker is a true nominal pronoun in an argument position and is resumptive of the left-dislocated topic with which it is coindexed (Zribi-Hertz & Diagne 2002:858-859).

As will be seen in the coming chapters, the syntactic status and underlying position of SCs are main points of contention in debates over SCs in many languages, including French, North Italian Dialects, Kreyòl Ayisen, and CVC; one side of the debate argues that SCs should be attributed the status that Zribi-Hertz & Diagne assigned to the atonic person markers in (16) (~'the inflectional affix hypothesis', cf. Culbertson 2010; Section 4.2, ahead) while the other side argues SCs are best analyzed along the lines of Zribi-Hertz' & Diagne's description for (17) (~'the phonological clitic hypothesis', cf. Culbertson 2010; Section 4.2, ahead).

Interestingly, and unlike what was observed in the descriptive accounts of CVC provided thus far, Wolof strong markers seem to violate some of the stipulations in Cardinaletti & Starke (1994, 1996, 1999) regarding the distributional and semantic-referential differences between strong, deficient, and clitic pronominals (see Section 4.3). For instance, Wolof strong subject markers can refer to animate or inanimate referents (18a-b), both clitic locative/object markers and a strong marker can be coordinated (19a-b), and both can receive interpretation as bound variable (20a-b) (Zribi-Hertz & Diagne 2002:861-862).

(18) a.) *{Moodu/ordinatër -bi},* *Aram bëgg -na* ***-ko***

Moodu/computer DET.DEF.SG Aram like +F 3SG.CL

'Moodu/the computer, Aram likes {him/it}.'

b.) *{Moodu/ordinatër -bi},* *Aram jendë -na buum ngir* ***moom***

Moodu/computer DET.DEF.S Aram buy +F wire for 3SG.STR

'Moodu/the computer, Aram has bought the wire for {him/it}.'

(19) a.) *(Moodu) Aram gis -na Nafi ak* ***moom***

Moodu Aram see +F Nafi and 3SG.STR

'(Mooodu) Aram sees both Nafi and him.'

b.) *(ordinatër -bi)* *Aram gis -na téeri -bi* *ak* ***moom***

computer DET.DEF.SG Aram see +F book DET.DEF.SG and 3SG.STR

'(The computer) Aram sees both the book and it.'

(20) a. *Aram$_z$ bëgg -na Moodu {gis* ***-ko*** */liggey ngir* ***moom$_z$****}, ak*

Aram want +F Moodu see 3S.CL /work for 3SG.STR and

Maryama$_k$ tam$_{(z/k)}$

Maryama too

'Aram$_z$ wants Moodu to {see her$_z$/work for her$_z$} and Maryama$_k$ too$_{(z/k)}$.'

b. *Ordinatër -bii lacc -na Moodu {gis* ***-ko$_z$*** */liggey ak* ***moom$_z$****},*

computer DEM need +F Moodu see 3SG.WK /work and 3SG.STR

ak téere -bii$_k$ tam

and book DEM too

'This computer$_z$, Moodu needs to {see it$_z$/work with it$_z$} and this book$_k$ too$_{(z/k)}$.'

Given the dual syntactic status of Wolof SCs/subject markers, the complex morphological transformations that they experience in different syntactic/pragmatic contexts, and their variable phonological attachment, Zribi-Hertz & Diagne (2002:862) reject outright the validity of a linguistic 'class' that singularly characterizes Wolof SCs/subject markers at the representational level of the grammar. Rather, they assert that:

> the phonological, syntactic, and morphological properties of Wolof person and locative clitics are independent from one another […]. From a phonological perspective, Wolof clitics exhibit one of the two regular behaviours of all weak-stress items in this language; from a syntactic perspective, they spell out either theta-marked expressions or inflected

person features; and from a morphological perspective, they instantiate weak words or word-level or phrase-level affixes.

With respect to the phonological properties of Wolof subject markers, they are no different from those of the other weak pronominals in the language (Zribi-Hertz & Diagne 2002:863-864). That is, weak functional items in Wolof cannot bear their own stress and undergo enclisis, except when they occur in initial position in the prosodic domain where they receive default stress, since Wolof is a stress-initial language. Thus, there is no phonological 'class' that can be attributed solely to clitics at the representational level, the relevant distinction in Wolof being between strong and weak items rather than clitics or non-clitics (Zribi-Hertz & Diagne 2002:864).

It was observed in (15a) that Wolof subject markers can occur in the same structural position as would a lexical subject, as a 'double-subject'-like construction (15b), or they can occur as person inflection markers, rather than as nominative arguments, as in (16a-c). In the case of the latter, it was concluded that subject markers emerge as person inflection markers because they occur in a "saturated finite clause" in which Tense, Aspect, and Polarity are fully spelled out. Zribi-Hertz & Diagne (2002:868-869) assert that this 'saturated finiteness' is akin to rich person inflection in consistent Null Subject Languages, and therefore in utterances like (21a-c) where subject marker *ñu* is analyzed as a person inflection marker, a lexical subject is possible, but not obligatory, in the phrase initial position (21a-b). This analysis implies that in (21b) there is a null subject identified by person inflection and thus the lexical subject cannot occupy the position of the person inflection marker (21c).

(21) a.) xale -yi lekk -na **ñu** ceeb -bi
child DET.DEF.PL eat +F 3PL rice DET.DEF.S
'The children have eaten the rice.'

b.) $ø_i$ lekk -na **ñu**$_i$ ceeb
3.PL eat +F 3PL rice
'They have eaten rice.'

c.) *lekk -na xale -yi ceeb

eat +F child DET.DEF.PL rice

Crucially, the syntactic status of Wolof subject markers, be it as SC pronouns or person inflection markers, holds independently of their phonological properties, which as was concluded, are sensitive to the strong/weak status of the item regardless of syntactic class.

In this subsection, evidence was advanced from Zribi-Hertz & Diagne (2002) that showed that the Wolof subject domain and the syntactic behavior of its pronominals bear many compelling resemblances, though also some crucial differences, to the CVC system. The fundamental generalization regarding the Wolof system is that phonologically, the relevant distinction is between weak and strong items, rather than [+/-clitic] items, and that the distributional and referential features associated with structural and referential deficiency in Cardinaletti & Starke (1994, 1996, 1999) do not always hold for Wolof pronominals. Further, Wolof atonic subject markers appear to 'lead a double life' as either SC pronouns or person inflection affixes. We will see in the coming chapters that such properties have been suggested for atonic SCs/person markers in many other languages. Further, this classification will become important for our conclusions regarding the status of SC in CVC at the end of this volume. I turn now to a discussion of the subject domain in Mandinka, proto-CVC's second most prominent substrate language.

3.3.2 Some aspects of the subject domain in Mandinka.

Three studies, Rowlands (1959), Davydov (2010), and Creissels (2015), provide details to establish a rudimentary understanding of the properties of the Mandinka subject domain. As was necessary with the Wolof data, given the paucity of historical data on Mandinka, I am forced to rely on accounts modern-day varieties for the discussion.

Just as with Wolof and several Atlantic languages, Manding languages, the group of which Mandinka forms a part, make use of emphatic or strong pronouns and non-emphatic weak/clitic

pronouns (Davydov 2010). According to Davydov (2010:23-24), doubling is possible in most contexts in Manding, though there are some contexts in which only emphatic pronouns are allowed, such as presentational contexts. The subject pronoun/marker inventory for Gambian Mandinka is shown in Table 19.

Table 19. Emphatic and non-emphatic subject pronouns in Gambian Mandinka (adapted from Rowlands 1959:55)

	Emphatic	Non-emphatic
1SG	*ńte*	*ńg, m, n*
2SG	*íte*	*í*
3SG	*àte*	*a*
1PL	*Ńtelu, ntolu*	*ńg, m, n*
2PL	*ál'telu, al'tolu*	*áli, al'*
3PL	*ìtelu, itolu*	*ì*

As with the Atlantic languages, Mande (of which Manding is a subgroup within which Mandinka is a language) subjects emerge in a position distinct from oblique or object nominals but display no overt (distinctive) subject marking (Creissels 2015:3). Unlike many Atlantic languages, there is no argument indexation for subjects nor are there null subjects even in anaphoric and expletive contexts, nor with a non-specific or arbitrary referent (Creissels 2010:2; 2015:6). The order of constituents in Mandinka clause structure is generally rigid, with a default order of S(O)V(X)(X') (where S=subject, O=object, V=verb. and X=oblique nominal) (Creissels 2010:2,5).

Unlike in Wolof in which 'strong person markers' can bear inanimate or animate reference, Rowlands (1959:55-56) observes that Mandinka unemphatic 3rd person singular *a* and third person plural *ì* can refer to either animate or inanimate referents, while their respective emphatic counterparts *àte* and *ìtelu* can only bear animate reference. He further observes that only emphatic

pronouns can occur in isolation. He also notes that, alternatively, demonstrative singular *wò* and plural *wòlu* can be used as emphatic subject pronouns with inanimate referents. We will see in the coming chapters that these distinctions are important components of typological models for classifying strong, weak, and clitic pronouns, such as that of Cardinaletti & Starke (1994, 1996, 1999).

With respect to the phonological properties of Mandinka clitics, Rowlands (1959:55) indicates that the forms in Table 17 "represent the pronunciation of these words when they occur initially in a sentence. In other positions they frequently fuse with a preceding word." Unemphatic forms undergo several transformations in "accent pattern" (tone), depending on if they are mono- or disyllabic, on their position in the phrase, on the intonational shape of adjacent elements, and based on combinations with functional items such as verb operators. Some of these transformations involve modifications in the surface morphophonological shape of the pronoun (Rowlands 1959:57-63). It is unclear if by "combination" and "fusion" the author is referring simply to modifications in intonation or tone, some sort of cliticization at the phonological level, or perhaps at the morphosyntactic level, or at all three levels.

For emphatic forms in phrase-initial position, "the accent patterns are similar to those of Nouns in this position", and it appears that they experience no 'fusion' processes, phonological or otherwise (Rowlands 1959:63-64). In a phrase-internal position, the emphatic forms can undergo variable 'fusion' with the constituent to the left: "Fusion of the first syllable [of the emphatic pronoun] occurs, though not so regularly as with Unemphatic and Reflexive Pronouns. Absence of fusion shows a more deliberate and emphatic pronunciation of the word" (Rowlands 1959:64). It is unclear if this is a purely phonological or prosodic process, or if the morphosyntactic context in which the pronoun occurs also determines the surface shape and intonational pattern of the pronoun. Indeed, the tonal and derivational properties of Mandinka subject pronouns have been more thoroughly studied than their morphosyntactic distribution or underlying structural properties. As a result, it is difficult to speculate as to the status of Mandinka unemphatic subject

pronominals with respect to whether their underlying morphosyntactic status may be more accurately described as inflectional person marker or genuine independent argumental pronominal.

The main contribution of this summary of the Mandinka subject pronoun system for the present study is that Mandinka, like Wolof, has a disjunctive tonic-atonic subject pronominal inventory. It appears that having these bipartite subject domains in both of its primary substrates was highly consequential for morphosyntactic restructuring in proto-CVC, since the subject pronoun inventory of proto-CVC developed an equivalent opposition in its own subject pronoun inventory. Since it is known that the development of a disjunctive tonic-atonic subject pronominal inventory was a common diachronic outcome in many Romance varieties (and in many languages with atonic subject pronominals), and that a similar opposition is operative among the superstrate's oblique pronominals, the emergence of the CVC bipartite system might be viewed as an instance of contact-driven cross-linguistic convergence (just as is likely the case for the 1st person SCs).

3.4 CHAPTER CONCLUDING REMARKS.

In this chapter, I reviewed a historical reconstruction of the proto-CVC subject pronoun inventory and its development into contemporary CVC (Lang 2012). I also summarized the previous literature on the organization of the subject domain and adjacent morphosyntactic properties in Late Medieval/Early Classical Portuguese, the superstrate language to proto-CVC on Santiago in the 15th and 16th centuries. Then, I consulted studies on subject pronominals in contemporary varieties of Wolof and Mandinka. While it is certain that there were many differences between modern and 15th-16th century varieties of these languages, we can nonetheless assume with some likelihood that the organization of the modern-day CVC overt subject pronoun system likely came about by cross-linguistic convergence, reflected in the observable disjunctive tonic-atonic opposition found in Wolof and Mandinka today, and perhaps an indication that older varieties of these languages that were contemporaneous to proto-CVC probably also had such oppositions in their subject pronoun inventories.

We can also conclude that proto-CVC underwent massive morphosyntactic restructuring in which it lost the following properties of its superstrate: consistent Null Subject Language status, suffixal subject-verb person-number agreement inflection (likely because the CVC verb is based

on the superstrate infinitive, cf. Quint 2008b), and the V2 reflex. The overt subject pronoun inventory of proto-CVC based its tonic pronominals on the morphophonological shape of the superstrate's tonic personal pronouns (either oblique or nominative, Quint 2001a), whereas various processes yielded the atonic SCs: phonological reduction of tonic forms, changes based on analogy with nominal plural inflection and other tonic-atonic parings in the system, and direct borrowing or imposition of two substrate atonic 'person markers' (1st person singular *N* and plural *nu*) (Lang 2012). Further, the development of a disjunctive tonic-atonic contrast in the CVC overt subject pronoun system is likely to also have arisen from a process of cross-linguistic convergence, since Wolof and Mandinka also feature such an opposition and since the development of such subject pronominal systems is known to be a commonly available outcome of diachronic change in Romance and other languages with atonic nominative pronouns. Finally, the observation that will be most relevant for the conclusion drawn in the present study for CVC SCs is with respect to Wolof 'subject markers'; these appear to function as both independent anaphoric pronouns as well as person inflection markers.

Chapter 4: An overview of previous research on subject clitics, null subjects/anaphoric zero, and double-subjects in Cabo-Verdean Creole and related languages.

This chapter is a review of the literature on the three anaphoric expressions under investigation in the present study: (i) subject clitics/atonic person markers; (ii) null subjects/zero anaphora; and (iii) double-subject constructions. This first half of the chapter is dedicated to examining the overt forms. I begin by examining two different classification of Cabo-Verdean Creole (CVC) overt pronouns, a tripartite classification based on the proposed classes of pronominals in Cardinaletti & Starke (1994, 1996, 1999), and a bipartite classification based on the disjunctive [+/-tonic] opposition.

This is followed by a discussion of two competing theories on the morphosyntactic status and hierarchical clausal positions of subject clitics that have repeatedly emerged in Generative Grammar-based research. These are the 'phonological clitic hypothesis' (cf. Kayne 1975; Déprez 1994; De Cat 2005; *inter alia*) and the 'inflectional affix hypothesis' (Brandi & Cordin 1989; DeGraff 1993; Baptista 1995, 2002; Poletto 2000; Culbertson 2010; *inter alia*). The former identifies subject clitics (SCs) as independent argumental pronominals in a specifier position that only cliticize at the phonological level. The latter argues for an analysis of SCs as bound inflectional affixes, syntactic heads in the IP/TP layer where they mark person-number concord.

An alternative theory, also developed in a Generative Grammar (GG henceforth) framework, is briefly addressed. Cardinaletti's & Starke's (1994, 1996, 1999) Typology of Structural Deficiency proposes three universally underlying classes of pronominal, strong, weak, and clitic; each class is associated with a set of morphosyntactic, prosodic, and semantic-referential properties.

I then consider research on SCs in disciplines outside the GG tradition. These approaches have traditionally used terms like subject marker, anaphoric person marker, or person agreement marker, among others, to describe roughly the same morphosyntactic element as SCs in the GG

tradition (cf. Siewierska 2004:121-126). Use of the terms subject marker usually implies something akin to 'phonological subject clitic', they are taken to be independent pronominal elements that engage in 'anaphoric agreement', a non-local anaphoric relationship with an antecedent outside of the immediate agreement field. Anaphoric person marker refers to a similar agreement relationship, but is perhaps indicative of the notion that the atonic anaphor has become bound or more dependent on its host. Person agreement markers seem roughly equivalent to assigning SCs the status of 'bound inflectional affixes'; they engage in 'grammatical agreement' with a local controller (Bresnan & Mchombo 1987), much like subject-verb person-number agreement marked on inflectional affixes as in consistent Null Subject Languages.

Further, brief consideration will be given to the role of SC/atonic person markers as "high referential continuity devices" in anaphora resolution (Givón 1976, 1983b, 2001[1984], 2017). This perspective considers 'referential continuity' in discourse organization to be the primary force determining the outcome of anaphora resolution; after zero anaphora, anaphoric atonic subject pronouns/markers are considered the referring expression for encoding the second highest level of referential continuity (cf. Givón 1976, 1983b, 2001[1984], 2017; Ariel 1990; *inter alia*). We will see that referential continuity is determined by a confluence of factors that facilitate 'antecedent accessibility', so that when an antecedent is at most accessible, activated, or salient in the discourse, anaphoric atonic subject markers become the second-most probable subject realization.

The second half of this chapter is dedicated to the issues of null subjects or zero anaphora. I begin with a review of the primary research program focused on this issue in the GG-tradition, the Null Subject Parameter. Special focus is reserved for partial Null Subject Languages, particularly those of Iberian-origin spoken in the Americas, as well as languages traditionally considered non-Null Subject Languages, such as English and creole languages. Since much of the research on subjects in CVC was conducted within the context of the Null Subject Parameter, I review studies like those of Baptista (1995, 2002), Pratas (2004), and Costa & Pratas (2008, 2013), which diverge on the issue of whether or not CVC is a partial or consistent Null Subject Language.

After exploring previous research on null subject/zero anaphora within the GG tradition, I then turn to research on this topic under 'Probabilistic Grammar'[32] approaches. I begin with investigations that were conducted on null subjects in Portuguese and Spanish using a variationist sociolinguistic approach. Then, I examine the closely related assumptions from functionalist, usage-based, 'cognitive linguistic', or typological approaches to zero anaphora.

To conclude this chapter, I summarize the discussion elaborated over its course. This serves to bring together assumptions from previous research on SCs and double-subject constructions with research on null subject/zero anaphora and what these entail for the exploration of these nominative anaphoric elements in Cabo-Verdean Creole in the present study.

4.1 THE CLASSIFICATION AND DISTRIBUTION OF OVERT SUBJECT PRONOUNS IN CABO-VERDEAN CREOLE.

CVC overt subject pronouns have been classified in two different ways, each reflecting the theoretical disposition of the researcher classifying them. Under the analyses that I will jointly call the 'bipartite analysis', the criterion for classification is word-level stress (i.e. whether the pronoun is tonic or atonic) (cf. Veiga 1995; Quint 2000a,b; Lang 2012). For those analyzing CVC from a GG perspective (e.g. Baptista 2002; Pratas 2004), CVC subject pronouns have been categorized according to the properties of the three classes postulated by Cardinaletti & Starke (1994, 1996, 1999) in their Typology of Structural Deficiency: strong, clitic, and weak. I will refer to this classification as the 'tripartite analysis'.

Manuel Veiga, in his descriptive grammar of CVC (1995:176-177,332-335), and his didactic workbook (2002:83), he does not extensively discuss the distribution of overt subject pronouns, nor their underlying structural properties, but does adopt the bipartite analysis in distinguishing atonic from tonic forms. Quint (2000b:161-163; 2015:44-45) also adopts the

[32] Recall that in the first chapter I adopted the term "Probabilistic Grammar" from Claes (2017) as a stand-in for various closely related sub-disciplines of linguistic inquiry such as functionalism, usage-based approaches, typology-oriented approaches, and corpus and variationist sociolinguistics. I will use this term, but sometimes switch among the others to elucidate several points later in the chapter.

bipartite analysis, but specifies that *a*-initial tonics forms must occur in utterance initial position, and with the notable exception of copular *é* 'to be', cannot immediately precede the verb phrase. He also observed that disyllabic (*a*-initial) tonic forms do not participate in focalizing constructions with relative complementizer *ki*, instead, only monosyllabic tonic forms are used (e.g. *(a*)mi ki ta manda* 'It is me (who is) in charge'). Another property of monosyllabic [+tonic] forms that is lacking in disyllabic [+tonic] forms is that the former occur after prepositions (*pa* ***mi*** 'for me', *ku* ***mi*** 'with me'), and with various adverbs such as *so* 'alone' (***mi****-so* 'me alone/just me'). Additionally, Quint notes that, while they are productive, *a*-initial forms are never required, and in all instances, they can be replaced by monosyllabic tonic forms (Table 20) (2000b:161-163; 2015:44-45). Another proponent of the bipartite analysis, Lang (2012:23) (see Table 12, Section 3.1), considers disyllabic tonic forms to be 'topic markers'.

Table 20. CVC [+tonic] subject pronouns categorized as "Tonic-initial" and "Tonic (simple)" with corresponding English translation (adapted from Quint 2000b:161-163)

Person-number	Tonic-initial	Tonic (simple)	English translation
1SG	*a-mi* /ɐ'mi/	*mi* /mi/	I
2SG	*a-bo* /ɐ'bo/	*bo* /bo/	you
2SG (polite, M)	*a-nhó* /ɐ'ɲɔ/	*nhó* /ɲɔ/	you (sir)
2SG (polite, F)	*a-nha* /ɐ'ɲɐ/	*nha* /ɲɐ/	you (madam/ma'am)
3SG	*a-el* /ɐ'el/	*el* /el/	he/she
1PL	*a-nós* /ɐ'nɔs/	*nós* /nɔs/	we
2PL	*a-nhós* /ɐ'ɲɔs/	*nhós* /ɲɔs/	you all
3PL	*a-ês* /ɐ'es/	*ês* /es/	they

Quint (2000b:164-165) refers to CVC subject clitics (SCs henceforth) as 'pre-verbal' ('*préverbal*') (Table 21[33]). He notes that only negator *ka* and TMA particles can intervene between a SC and the verb, and in absence of these intervening elements, the SC attaches to the verb (when not attaching to some other host, see 31-33 ahead). The 1st person singular SC, often transcribed as

[33] Examples are given with non-stative bare verb *kanta* 'sing' rendering a default [+past] reading.

N, is a nasal element or "trace of nasality" that induces nasalization of the following consonant or the preceding vowel (cf. Brito 1867[1887], cited in Lang 2012:27). With respect to the availability of both *el* and *e* in the 3rd person singular, Quint considers this to be the result of a change in progress whereby the more conservative form *el* is being lost to the variant with the lone vowel *e* in younger generations (2000b:164-165).

Table 21. CVC [-tonic] 'pre-verbal' subject pronouns (SCs) (adapted from Quint 2000b:164-165)

Person-number	Atonic form	Example	Translation
1SG	*m-* /~/	*m-kanta* /'~kɐ̃tɐ/	I sang
2SG	*bu* /bu/	*bu kanta* /bu'kɐ̃tɐ/	you sang
2SG (polite, M)	*nhu* /ɲu/	*nhu kanta* /ɲu'kɐ̃tɐ/	you sang (sir)
2SG (polite, F)	*nha* /ɲa/	*nha kanta* /ɲa'kɐ̃tɐ/	you sang (mam)
3SG	*el/e* /el, e/	*el/e kanta* /el'kɐ̃tɐ, e'kɐ̃tɐ/	he/she sang
1PL	*nu* /nu/	*nu kanta* /nu'kɐ̃tɐ/	we sang
2PL	*nhós* /ɲɔs/	*nhós kanta* /ɲɔs'kɐ̃tɐ/	you all sang
3PL	*ês* /es/	*ês kanta* /es'kɐ̃tɐ/	they sang

Quint (2000b:165) also affirms that SCs can attach to an immediately preceding complementizer such as *pa, ma,* or *ki* as pre-verbal enclitics (Table 22). In addition, Lang (2012:22$_{n4}$) notes that the same type of enclitic attachment occurs with presentational particles *alâ-**l** ta ben!* 'there he/she comes', the adverb 'already' or perfective TMA particle *dja - si dja-**u** kre bai,* 'if you want to go already', negator *ka - ka-**u** mata-m!* 'don't you kill me!', and even to a tonic pronoun itself in an amalgamated construction *abo-**u** ta fika li!* 'you, you'll stay here!'.

Table 22. Cabo-Verdean Creole pre-verbal enclitic subject pronouns (adapted from Quint 2000b)

Person-number	Atonic form	Examples: *e kre…*	Translation: *I want…*
1SG	*-m* /~/	*pa-m kanta* /pã'kɐ̃tɐ/	for/Ø me/Ø to sing
2SG	*-u* /w, u/	*pa-u kanta* /pɔw'kɐ̃tɐ/	for/Ø you to sing
2SG (polite, M)	*nhu* /ɲu/	*pa nhu kanta* /pɐ ɲu'kɐ̃tɐ/	for/Ø you to sing (sir)
2SG (polite, F)	*nha* /ɲa/	*pa nha kanta* /pɐ ɲa'kɐ̃tɐ/	for/Ø you to sing (ma'am)
3SG	*-e* /e/	*p-e kanta* /pe'kɐ̃tɐ/	for/Ø he/she to sing
1PL	*-nu* '-/nu/	*pá-nu kanta* /'panu'kɐ̃tɐ/	for/Ø us to sing
2PL	*nhós* /ɲɔs/	*pa nhós kanta* /pɐ ɲɔs'kɐ̃tɐ/	for/Ø for you all to sing
3PL	*-ês* /es/	*p-ês kanta* /pes'kɐ̃tɐ/	for/Ø them to sing

The tripartite analysis, on the other hand, is adopted by researchers working within a GG framework such as Baptista (2002) and Pratas (2004). It is rooted in the Typology of Structural Deficiency (Cardinaletti & Starke 1994, 1996, 1999) and adopts the view that CVC subject pronouns correspond to the three universally underlying categories of pronominal: strong, weak, and clitic. The formal features that are hypothesized for each category are discussed in more detail in the next subsection. For the time being, I note that tonic disyllabic forms are classified as strong, monosyllabic forms as weak, and atonic forms as SCs (Table 23).

Table 23. The tripartite classification of overt subject pronouns in CVC (adapted from Pratas 2002:42)

Person-number	Strong	Weak	Clitic
1SG	*ami*	*mi*	*N*
2SG	*abo*	*bo*	*bu*
2SG (polite, M)	*anho*[34]	*nho*[35]	*nhu*
2SG (polite, F)	*anha*	*nha*	*nha*
3SG	*ael*	*el*	*e*
1PL	*anos*	*nos*	*nu*
2PL	*anhos*	*nhos*	*nhos*
3PL	*aes*	*es*	*es*

SCs can occur as the sole subject element, as can a weak or strong pronoun, or SCs can double with a weak or strong form in a double-subject construction (X2SBJ henceforth) in which the coreferential clitic and strong pronoun are amalgamated (adapted from Pratas 2004:43) Table 24)[36].

34 Pratas records this form as *anhu*.

35 Pratas records this form as *nhu.*

36 X2SBJs in Pratas (2004:43) are recorded with the weak or strong pronominal separated from the clitic by an orthographic comma, presumably implying an intonational pause. This is because she adopts the view that SCs are in the subject position, and therefore lexical DPs and tonic pronouns doubling the SC must necessarily be left-dislocated (see Section 4.2, ahead). I noted above, however, that there is often no pause between the strong/weak pronoun and its coreferential SC; indeed, in the first, second, and third person singular, the SC often attaches directly to the tonic form (see also Lang 2012:22$_{n4}$; Baptista 2002:50,256): *ami=N kanta, abo=u kanta, ael=e kanta.*

Table 24. Double-subject constructions in Cabo-Verdean Creole

Person-number	Double-subject constructions	Translation:
1SG	*(a)mi N kanta*	'I sang'
2SG	*(a)bo bu kanta*	'You sang'
2SG (polite, M)	*(a)nho nhu kanta*	'You sang (sir)'
2SG (polite, F)	*(a)nha nha kanta*	'You sang (ma'am)'
3SG	*(a)el e(l) kanta*	'He/she sang'
1PL	*(a)nos nu kanta*	'We sang'
2PL	*(a)nhos kanta*	'You all sang'
3PL	*p-ês kanta* /pes'kɐ̃tɐ/	'They sang'

Pratas (2004) examines the distribution of CVC subject pronouns by applying a series of syntactic tests associated with some of the asymmetries from Cardinaletti & Starke (1994, 1996, 1999), which in turn are partially based on Kayne's (1975:81-85) analysis of French clitics. To begin with, Pratas confirms that scs cannot be separated from the verb by any intervening constituents, such as adverbials, with the exception of TMA particles and negator *ka* (22a-e; adapted from Pratas 2004:48).

(22) a.) *E abri porta*

3SG open.PAST.PERF door

'He opened the door'

b.) *E ta abri porta*

3SG TMA.HAB open door

'He opens the door.' (>always/usually/frequently)

c. *E ka abri porta*

3SG NEG open.PAST.PERF door

'He did not open the door'

d. *E ka ta abri porta*

3SG NEG TMA.HAB open door

'He does not open the door'

e. **E senpri abri porta*

3SG always open door

'He always opens the door'

Additionally, she asserts that SCs cannot be coordinated with another DP (23a-b) and cannot be modified (24a-b), and cannot be focalized (25a), nor can they occur in isolation (25b) (adapted from Pratas 2004:47-55).

(23) a.) **N ku Maria ta bai beiramar*

CL and Maria TMA go beach

b.) **Maria ku N ta bai beiramar*

Maria and CL TMA go beach

'Maria and I go to the beach.'

(24) a.) **Nu dos ta ganha txeu dinheru*

CL two TMA earn a lot money

'We two earn a lot of money'

b.) **E* so ta ganha txeu dinheru

CL alone TMA earn a lot money

'He earns a lot of money on his own/by himself'

(25) *Kenhe ki ganha txeu dinheru?*
'Who earns a lot of money?'

a.) **E* ganha txeu dinheru
CL earn a lot money

b.) **N*.
CL

In contrast with SCs, she proposes that CVC weak subject pronouns can be coordinated with another DP (26a-b) and can be modified (27a-c), but like SCs they cannot be focalized (28a-b), nor occur in isolation (29a-c[37]) (Pratas 2004:52-55).

(26) a.) *Mi ku Maria, nu ta bai beiramar*
WK.1S CONJ M. CL.1P TMA go beach

b.) *Maria ku mi, nu ta bai beiramar*
M. CONJ WK.1S CL.1P TMA go beach
'Maria and I, we go to the beach'

(27) a.) *Nos dos ta ganha txeu dinheru*
WK.1P two TMA earn a lot money
'We two earn a lot of money'

b.) *El so ta ganha txeu dinheru*
WK.3S alone TMA earn a lot money
'He makes a lot of money on his own/by himself'

[37] 29b is claimed to be subject to dialectal variation.

(28) *Kenhe ki ganha txeu dinheru?*

'Who earns a lot of money?'

a.) **El ganha txeu dinheru*

WK.3S earn a lot money

b.) *Ael, e ganha txeu dinheru*

STR.3S CL.3S earn a lot money

'HE, he earns a lot of money'

(29) *Kenhe ki ta ganha txeu dinheru?*

a.) **N*

CL.1SG

b.) ** Mi*

WK.1 SG

c.) *Ami*

STR.1SG

The copular verb *é* 'to be', including its corresponding irregular forms inflected for aspect *era* and *foi*, is the only verb that disallows SCs (30) (Pratas 2004:55-56).

(30) a.) *Mi e/era spertu*

WK.1S COP expert

b.) **N e/era spertu*

CL.1S COP expert

'I am/was an clever.' (~sly, smart)

In cases of SC complementizer attachment where the SC is 3rd person and coreferential with the matrix subject, the surface form of the attached clitic can be that of accusative clitic *-l* or nominative clitic *-e* (31a-b). In cases where the embedded subject is not co-referential with the matrix subject, some of Pratas' (2004:56-57) informants find the surface accusative form "strange" (31b).

(31) a.) *Pedru$_i$ fri p'-**e**$_i$ podi falta na trabadju*
Pedro injure COMP-**CL.NOM.3S** could miss in work
b.) *Pedru$_i$ fri pa-**l**$_i$ podi falta na trabadju*
Pedro injure COMP-**CL.ACC.3S** could miss in work
'Pedro injured himself so (that) he could miss work.'

(32) a.) *Djon$_i$ fla-l p'-**e**$_j$ bai si kaza.*
John say-CL.ACC.3S COMP-**CL.ACC.3SG** go POS.3SG house
b.) *?Djon$_i$ fla-l pa-'**l**$_j$ bai si kaza.*
John say-CL.ACC.3S COMP-**CL.ACC.3S** go POS.3SG house
'John said for him/her to go to his/her house'

In all other grammatical persons, the surface form of the clitic can be the accusative even in cases where there is no coreference between the matrix and embedded subject (33a-d) (Pratas 2004:57).

(33) a.) *Djon$_i$ fla pa-**m**$_j$ bai si kaza*
John say COMP-**CL.ACC.1SG** go POSS.3S house
'John told me to go to his house'

b.) *N_i* *fri* *pa-**m**$_i$* *podi falta trabadju*

CL.NOM.1S injure COMP-**CL.NOM.1SG** could miss work

'I injured myself so I could miss work'

c.) *Djon$_i$ fla pa-**u**$_j$* *bai si* *kaza.*

John said COMP-**CL.ACC.2SG** go POS.3SG house

'John said for you to go to his house.'

d.) *Bu$_i$* *fri* *pa-**u**$_i$* *podi falta trabadju.*

CL.NOM.2S injure COMP-**CL.NOM.2SG** could miss work

'You injured [yourself] so you could miss work.'

Baptista (2002:213-250), in her exposition on subject pronouns in the *Sotavento* varieties of CVC, also dedicates a significant portion of her analysis to demonstrate that the framework in Cardinaletti & Starke (1994, 1996, 1999) describes well the syntactic distribution of SCs). However, she does differ slightly from Pratas (2004) in some of her other claims (not enumerated here) regarding the specific distributional properties of CVC subject pronouns. Further, despite theoretically adopting the tripartite analysis, she nevertheless tends to present the subject pronoun inventory according a bipartite categorization.

As was noted in Chapter 1, these classic syntactic/distributional 'tests' for clitichood have proven unable to resolve the protracted pronoun vs. affix debate, hence the motivation for applying a variationist sociolinguistic methodology in the present study. A comprehensive application of the full set of classic tests for clitichood like those of Kayne (1975), Klavans (1982), Zwicky & Pullum (1983), and Cardinaletti & Starke (1994, 1996, 1999), against the distribution of CVC subject pronominals in my corpus, is left for a future project. Nevertheless, this subsection has provided a base description of the basic morphological shape and syntactic distribution of CVC overt pronouns.

4.2 SUBJECT CLITICS IN ROMANCE: PHONOLOGICAL CLITICS OR INFLECTIONAL AFFIXES?

In this subsection, I turn to research on X2SBJs and SCs in Romance languages and Romance-lexifier creoles. This discussion is kept separate from the discussion of typological investigations into 'person markers', which is the term that is typically used for SC-like morphemes in non-Indo-European languages. Despite many similarities, there are notable differences in the terminological and theoretical assumptions in research on SC in Romance and Romance-adjacent languages, as opposed to research on person markers in non-Indo-European languages. In order to avoid confusion, I deal with research into person markers in the following subsection. Before that, it will be useful to consider disagreements over the status of SCs in Romance and Romance-based creoles, since the debate over SCs in CVC was developed along the same lines.

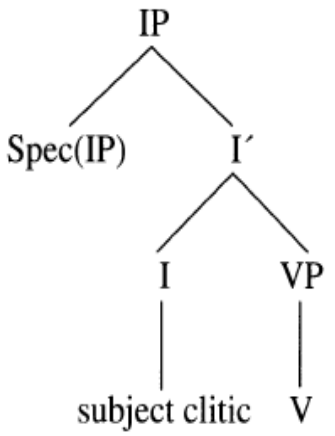
X2SBJs and SCs have been explored in several Northern Italian Dialects (cf. Renzi & Vanelli 1983; Benincà 1984; Rizzi 1986b; Brandi & Cordin 1989; Poletto 2000; Goria 2004; *inter alia*), Gallo-Romance varieties including Franco-Provençal and Occitan (Savoia & Manzini 2010; Benincà 2014), Swiss French (Fonseca-Greber 2000), Québécois (Roberge 1990; Auger 1994a), and Metropolitan French (Lambrecht 1981; Zribi-Hertz 1994; De Cat 2005; Culbertson 2010; Palasis 2015; *inter alia*). Though none of these were source-languages for CVC[38], nor are they directly related to CVC, they are relevant examples of disjunctive tonic-atonic subject pronoun systems in Romance. Further, disagreements over the morphosyntactic status and clausal position of SCs in these languages are reflected in parallel disputes over these elements in Romance-based creole languages like Kreyòl Ayisen (DeGraff 1993, Déprez 1994) and CVC (Baptista 1995, 2002; Pratas 2004).

[38] Although Genovese and Venetian mariners, traders, and settlers were notably present in the early phase of expansion of the Kingdom of Portugal's maritime empire and during the settlement of Santiago, there is, to the best of my knowledge, no evidence of sufficient numbers of settlers from these regions to entertain the possibility that these languages might have played a prominent role in the linguistic 'feature pool' during the early formative years of proto-CVC.

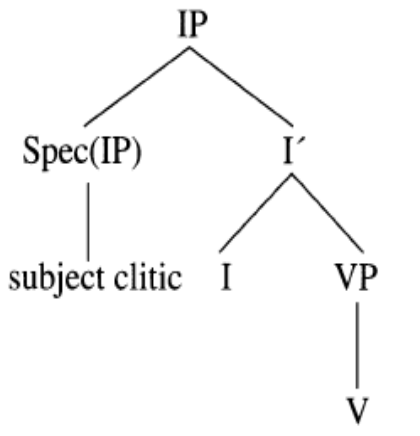
The theoretical dispute over the morphosyntactic status and clausal position of SCs in these languages was characterized by Culbertson (2010) as consisting of two hypotheses that have been revisited by researchers exploring SCs in many different languages: (i) the 'inflectional affix hypothesis' and (ii) the 'phonological clitic hypothesis'. The inflectional affix hypothesis states that SCs are bound inflectional affixes in a head position within the IP/TP layer (34a), while the phonological clitic hypothesis claims that SCs are independent pronouns occupying a specifier position (35b).

(34) Competing hypothesis for the position of SCs in Romance (Culbertson 2010:89)

a. Inflectional affix hypothesis



b. Phonological clitic hypothesis



These competing hypotheses were originally used to differentiate SCs in Northern Italian Dialects from SCs in French. The former has also been adopted for SCs in Kreyòl Ayisen (DeGraff 1993) and in CVC (Baptista 1995, 2002), while the latter has been adopted *en contra* these views by Déprez (1994) for Kreyòl Ayisen, and by Pratas (2002) and Costa & Pratas (2008, 2013) for CVC.

Some of the evidence proposed for the differential status of SCs in the Northern Italian Dialects and French have come from the obligatoriness of SCs in certain contexts. In Fiorentino and Trentino, SCs are obligatory in certain person-number configurations (35a-b), with a lexical

DP (36), a tonic subject pronoun (37) (Poletto 2000b; Palasis 2015), and even an indefinite quantifier (38) (Brandi & Cordin 1989:111; Culbertson 2010:86).

(35) a.) **Parli* (Fiorentino, Trentino)
speak

b.) *Tu parli* (Fiorentino)
Te parli (Trentino)
2SG.CL speak
'You speak.'

(36) *Mario *(e) parla.* (Fiorentino)
M. 3SG.CL speak
'Mario, he speaks.'

(37) *Ti te magni sempre.* (Vèneto)
2SS.STR 2SG.CL eat always
'You always eat'

(38) *Nessuno *(gl)=ha detto nulla.* (Fiorentino)
Nobody 3.SG.CL.INDEF=AUX say nothing
'Nobody said anything'

Contrast this with French where the SC is presumably optional with a definite subject DP (39) and disallowed with an indefinite quantifier (40) (Culbertson 2010:86).

(39) *Jean (il) parle.*

J. 3SG.CL speak

'John speaks.'

(40) *Personne (*il) n'a rien dit.*

Someone 3.SG.CL NEG-AUX nothing say

'Nobody said anything.'

Thus, in apparent doubling contexts, the French DP is taken to be left-dislocated in a topic position while the SC is in SPEC,IP (41a-b) (Rizzi 1997), and doubling with an indefinite DP or operator should be disallowed since topics are associated with definiteness (41c-d) (Palasis 2015:126).

(41) a.) [$_{IP}$ *il parle*]

3SG.CL speak

'he speaks'

b.) [$_{TOPP}$ *l'* *enfant* [$_{IP}$ *il* *parle*]]

DET.DEF child 3SG.CL speak

'the child speaks'

c.) *[$_{TOPP}$ *un* *enfant* [$_{IP}$ *il* *parle*]]

DET.INDF child 3SG.CL speak

'a child speaks'

d.) *[$_{TOPP}$ *quelqu'un* [$_{IP}$ *il* *parle*]]

someone 3SG.CL speak

'someone speaks'

Several other syntactic tests have been used to demonstrate the different underlying positions of SCs in French and Northern Italian Dialects; among these are the position of SCs relative to negation, the repetition of SCs in VP-coordination structures, and SC-verb inversion in interrogatives contexts, among others (cf. Kayne 1975; Rizzi 1986a; Brandi & Cordin 1989; *inter alia*).

Many of these claims have not been found to hold across Northern Italian Dialects or can be accounted for by other structural considerations (cf. Poletto 2000). For example, Zanutti (1997) and Poletto & Totora (2016:774) point out that differences in the position of SCs with respect to negation may be related to different underlying positions for negation in the clause, particularly in varieties with more than one available position for a negator, rather than the underlying position of the SCs themselves[39]. Similarly, in some Northern Italian Dialects, coordinated VPs will obligatorily repeat SCs across clauses, whereas other Northern Italian Dialects follow the French pattern of not requiring repetition of the SCs across coordinated VPs, and those varieties that do require repetition of the SC employ a particular class of SCs that adheres to adjacency requirements relative to the verb (Poletto & Totora 2016:774)[40]. French maintains variable SC-verb inversion in interrogatives, a feature that was lost in some Northern Italian Dialects but is maintained in others (Pescarini 2018). Clearly, these classic tests of clitichood have proven incapable of uncovering the morphosyntactic status and underlying position of SCs across Northern Italian Dialects[41].

What is more, there is a robust body of research showing that the status of SCs in Gallo-Romance may be related to issues of register or dialect variation, such that the morphosyntactic status and underlying clausal position of SCs in Standard Metropolitan French may differ from that of SCs in Colloquial Metropolitan French and regional varieties of French such as Québécois, Ontario French, and Swiss French (cf. Lambrecht 1981; Roberge 1990; Auger 1994a; Fonseca-

39 See the discussion ahead concerning Culbertson (2010:93-99) on the position of negation and SCs in Colloquial Metropolitan French.

40 See also the discussion ahead (Section 4.5), from Torres Cacoullos & Travis (2019), who refute the validity of VP-coordination tests for identifying the status of pronominals based on quantitative analysis of naturalistic English data.

41 See Poletto (2000b) and Goria (2004) for cross-dialectal surveys of variable forms of SCs in Nothern Italian Dialects that evade simplistic syntactic classification.

Greber 2000; Culbertson 2010; Palasis 2015; *inter alia*). These differences lie in the syntactic distribution of SCs, their morphophonological properties, as well as the frequency and status of double-subjects, and the range of subject elements available for doubling with SCs (Culbertson 2010, and references therein; Palasis 2015, and references therein). The claim is that accounts such as Kayne (1975), Rizzi (1986b), Brandi & Cordin (1989), and De Cat (2005) adequately describe the distribution of SCs in Standard Metropolitan French, but not in Colloquial Metropolitan French or regional French vernaculars.

Culbertson (2010) synthesizes the literature since the 1970s on SCs and X2SBJs in Gallo-Romance and combines this previous research with her own results drawn from various corpora and an auditory grammaticality judgement task, among other sources. She concludes based on a combination of phonological, syntactic, and typological evidence, that SCs in Colloquial Metropolitan French have experienced grammaticalization resulting in erosion to their pronominal status, making them more affix-like.

Generally, Colloquial Metropolitan French SCs are more phonologically reduced than in Standard Metropolitan French, an observation that Culbertson (2010:91) takes to be one sign of the grammaticalization process, since it is well known that phonological reduction resulting from grammaticalization can also be indicative of loss of morphosyntactic independence and attachment to an adjacent host (cf. Hopper & Traugott 1993:7). Syntactic evidence that Colloquial Metropolitan French SCs are inflectional affixes comes from distributional differences between Colloquial Metropolitan French and Standard Metropolitan French in the position of negation relative to the SC and the verb, and differences in the use of SCs in VP-coordination structures.

For negation, the intervention of negator *ne* between the verb and SCs was taken as evidence that SCs are nominative arguments in a specifier position (Rizzi 1986a; Brandi & Cordin 1989). However, in Standard Metropolitan French, negation is discontinuous; the first negator *ne* appears pre-verbally between the verb and the subject position, while a second negator *pas* appears post-verbally. Culbertson (2010:93-95), following prior studies (Ashby 1982; Armstrong 2002; Coveney 2002; *inter alia*), showed that *ne*-retention occurs at very low rates in Colloquial

Metropolitan French, leaving only post-verbal negator *pas* and no negator intervening between the SC and the verb. Rates of *ne*-retention are also sensitive to the prior subject type. In a measure of rates of *ne* negation in corpus of Colloquial French from Lyon, Culbertson (2010:95) found that *ne* was retained most often with a lexical DP subject (83.3%), second most often with constructions like imperatives and impersonals that employ neither a lexical DP nor a SC (14.1%), third most often with lexical DP + SC double-subject constructions (6.7%), and least often with singleton SC subjects (6.3%). Culbertson (2010:99) argues that the low overall rate of *ne*-retention in Colloquial French and the tendency to retain *ne* with subject types that do not contain SCs is indicative of the increasingly affixal status of SCs in this register. She further argues that the residual use of *ne* between SCs and the verb can be attributed to speakers switching between their Colloquial Metropolitan French 'grammar', in which SCs are inflectional affixes, and their Standard Metropolitan French 'grammar', in which SCs are independent subject pronouns. In CVC, the primary negator is *ka*, which as we have seen intervenes between SCs and the verb (22c-d, above). This may have more to do, however, with the underlying position of negation rather than the SC itself (Zanutti 1997; Poletto & Totora 2016:774).

In Standard Metropolitan French, subjects variably invert with the verb in WH-questions and yes-no questions. Inversion of SCs and the verb in these contexts might be taken to be indicative of the non-affixal status of SCs. However, in several corpora of Colloquial French speech, SC-verb inversion is fleeting (Culbertson 2010:100-101); Coveney (2002) and De Cat (2007) find no instances SC-verb inversion in yes-no questions, and only 6.6% and 2% inversion in WH-questions, respectively. In Culbertson's Lyon corpus, she examined various subject types for subject-verb inversion; she found that lexical DPs inverted with the verb at a rate of 2.4% in yes-questions and 88.5% in WH-questions, singleton SCs inverted with the verb at a rate of 0.1% in yes-no questions and at a rate of 0.9% in WH-questions, and lexical DP + SC double-subjects never inverted in either context (2010:100-101).

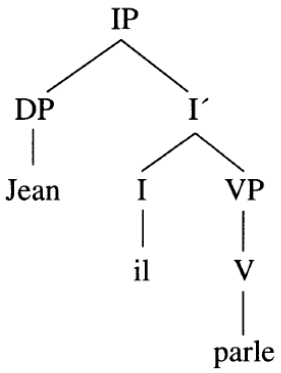
Just as she concluded for negation, Culbertson (2010:101) takes the data from inversion structures as evidence in favor an affixal analysis of Colloquial French SCs. She arrives upon

parallel conclusions for VP-coordination structures as well; the Lyon corpus revealed rates of repetition of SCs across coordinated VPs >95% in all person-number instantiations, lending further support to affixal analysis for Colloquial French SCs (Culbertson 2010:102-103).

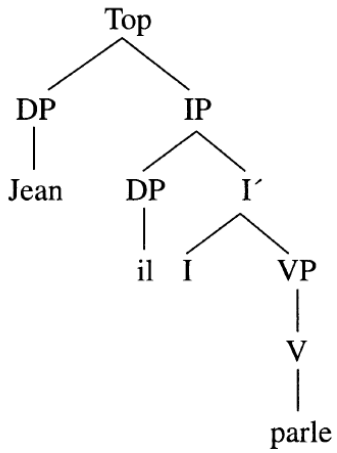
Two additional clues from Colloquial French X2SBJs point to the affixal status of SCs. Recalling the tree structures for the inflectional affix hypothesis (34a, above) and the phonological clitic hypothesis (34b, above), consider the implications for X2SBJs: in the case of double-subjects constructed from a lexical DP and an SC (rather than a tonic pronoun + a SC), the phonological clitic hypothesis assumes that the lexical DP is left-dislocated, since SCs are in the canonical subject position (SPEC,IP) (43a). On the other hand, the inflectional affix hypothesis assumes that lexical DPs occupy the subject position in X2SBJs, since SCs are inflection heads in VP layer, yielding a "true subject-doubling structure" (43b) (Culbertson 2010:104-105).

(42) Competing hypotheses for double-subjects in French (Culbertson 2010:104)

a. Left-dislocation structure



b. Subject-doubling structure



Diachronic and typological studies on subject doubling structures have found that increased reliance on left-dislocated lexical DPs can induce grammaticalization of atonic subject pronominals ('the NP-Detachment Hypothesis'), driving them towards bound inflectional affix status in what is known to be a cross-linguistically common grammaticalization cline (cf. Givón

1976; Hopper & Traugott 1993:7; Siewierska 2004). As pronominal SCs become more affixal over time, X2SBJs ("true doubling structures") should become more frequent, and the lexical DPs in the X2SBJs should show less signs of left-dislocation. Culbertson (2010:105-123) provides extensive evidence that this is occurring with Colloquial French X2SBJs. One piece of evidence comes from the prosodic contour of the lexical DP in X2SBJs and the another comes from overall rates of X2SBJs.

In an experimental auditory study, Culbertson examined the prosodic contour of lexical DPs in X2SBJs by comparing them to other unambiguously left-dislocated elements (2010:106-111). She found that lexical DPs in X2BJs no longer bear the prosodic features of other left-dislocated elements, suggesting that lexical DPs are not left-dislocated when doubled with SCs, and further implying that the SCs have affixal status. As was seen in Table 22 (and footnote 34), this also true of X2SBJs involving a tonic pronoun and SC amalgamation in CVC, in that there is often no audible pause between the two elements, in fact, there is often direct attachment of the SC to the tonic form in the singular forms.

Culbertson also considered overall rates of X2SBJs in oral corpora of adult and child-directed speech (2010:115-117). Overall rates of X2SBJs with a lexical DP and a SC vary greatly depending on the variety, register, and methodology employed for data collection. Some previous studies have found rates of X2SBJs involving a lexical DP around 20%-35% in adult-directed French (Ashby 1980; Nadasdi 1995; Coveney 2002), but as high as 60% in the in the *Phonologie du français contemporaine* corpus (Culbertson 2010:90,116). In this latter corpus, there are differences for age, with those under 35 years old using X2SBJs up to 70% of the time and those older than 35 using them 50% of the time (Culbertson 2010:116). In child-directed speech, Culbertson finds that rates of X2SBJs with lexical DP + SC climb to 81%. Culbertson argues that the difference in rates of subject doubling in child-directed speech as opposed to adult-directed speech are again indicative that register selection is the primary mechanism determining the occurrence of X2SBJs in spontaneous speech (2010:116-117).

Based on the preponderance of the evidence that SCs in Colloquial Metropolitan French have affixal status, Culbertson (2010:121) proposed a typology for X2SBJs across varieties of Northern Italian Dialects and Gallo-Romance; this typology captures the range possible of subject constituents available for doubling (Table 25). Given the independent subject pronoun status of scs in Standard French, "true doubling" is considered impossible in all contexts, since lexical DPs, tonic subject pronouns, and other subject constituents are always considered left-dislocated when an SC is present. In Colloquial Metropolitan French, on other hand, SCs are tense heads, so strong pronouns and lexical DPs can double the SC, presumably appearing in subject position; however, doubling with indefinite subjects, quantifiers, and other operators such as WH-variables is much more restricted, as can be observed in the low rates of doubling in these contexts in Culbertson's data (2010:122-114). Finally, languages like Fiorentino and the Gallo-Romance vernacular Picard allow productive doubling with various subject constituents.

Table 25. A typology of subject doubling across varieties of Northern Italian Dialects and French (adapted from Culbertson 2010:121)

Language	Pronoun	Definite DP	QP	WH-variable
Standard French	-	-	-	-
Central Veneto	+	-	-	-
Colloquial French	+	+	-	-
Trentino	+	+	+	-
Fiorentino/Picard	+	+	+	+

It is possible that these cross-Romance differences in the elements available for doubling with SCs are indicative of the degree of grammaticalization on the SC on its diachronic path towards inflectional affix. Certainly, Culbertson's results have demonstrated that, at least in colloquial registers of French, SCs have experienced some degree of degradation in their pronominal status

relative to the more conservative standard register in which they maintain their fully pronominal status.

Crucially, this transformation is likely a continuation of a much older change in the history of the French: Old French was a consistent Null Subject Language with 'rich' inflectional morphology marking subject-verb person-number agreement, but lost its Null Subject Language properties as the inflectional paradigm eroded through Middle French and into Modern French (Roberts 1993:125, Kaiser 2009). It is possible that this loss of inflection also played a role in inducing the grammaticalization of pronominal SCs into more affixal SCs (cf. Wratil 2011), a point to which I will return later in this subsection. Culbertson's study has shown that synchronic data are capable of producing evidence of the diachronic erosion of independent pronominal SCs into affix-like SCs, and especially with respect to X2SBJ constructions.

Interestingly, research into Brazilian Portuguese has also shown that X2SBJ constructions have become more frequent in that language (44a-d) (Kato 1999; see also Barbosa, Duarte, & Kato 2005, 2008; Camacho 2008, 2016; Tavares Silva, Carvalho, & Ziober 2018).

(43) a.) *[A Clarinha]$_i$ ela$_i$ cozinha que é uma maravilha.*
'Clarinha, she cooks wonderfully.'

b.) *Então [o Instituto de Física]$_i$ ele$_i$ manda os piores professores… [Os melhores]$_i$ eles$_i$ dão aula no curso de matemática*
'Then the Institute of Physics it sends the worst professors... The best, they teach in mathematics.'

c.) *Eu, eu sinto demais isso.*
'Me, I feel that too much.'

This strategy has become such a prominent feature of discourse organization in Brazilian Portuguese that some have argued for classifying it as a topic-comment or topic-prominent

language[42] (Pontes 1987; Negrão & Viotti 2000; Modesto 2008; Oliveira da Silva & Álvez Fonseca 2018). With respect to the 'typology' of x2SBJs in Table 22 (above), and the semantic properties of the participating lexical DP, Tavares Silva, Carvalho, & Ziober (2018) have shown that Brazilian Portuguese allows for X2SBJs with indefinite lexical DPs, albeit at very low rates.

According to the 'NP-Detachment Hypothesis' (Givón 1976; Siewierska 2004:263,264), an historical increase in rates of X2SBJs should be indicative of cliticization of the independent pronoun involved in the amalgamated structure. Indeed Vitral (1996), Ramos (1997), and Luís & Kaiser (2016:230), have all argued for the increasingly SC-like status of Brazilian Portuguese subject pronominals (Table 26).

Table 26. Reduced second- and third-person subject pronouns in Colloquial Brazilian Portuguese (adapted from Luís & Kaiser 2016:230)

	Standard form	Reduced form
2SG	*você*	*cê*
3SG (M)	*ele*	*el*
3SG (F)	*ela*	*é*
2PL	*vocês*	*cês*
3PL	*eles*	*es*

Notice that the third person forms are nearly homophonous to CVC third person SCs[43]. In a parallel acknowledgement of the grammaticalization process acting on Brazilian Portuguese subject pronouns, Kato (1999) analyzes Brazilian Portuguese subject pronouns as "[+pronominal] AGR". In other words, they have lost some of the properties usually associated with nominative pronominals, and to a certain extent, have begun to exhibit features of agreement markers.

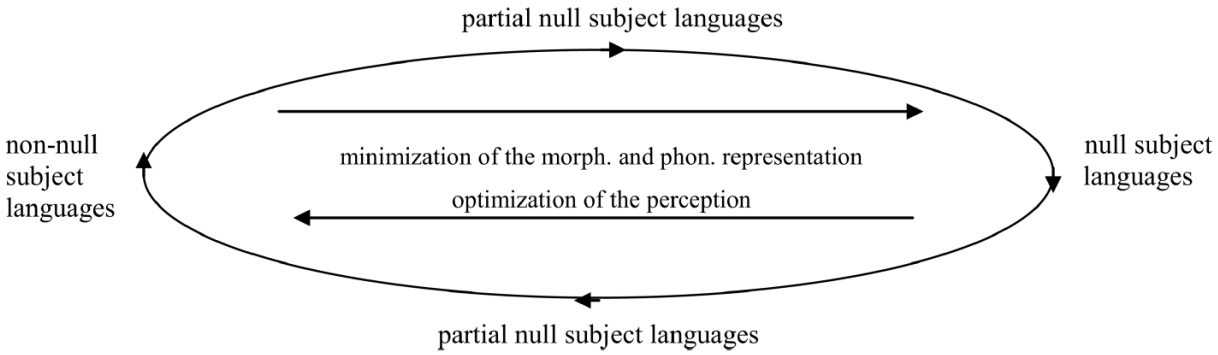
Wratil (2011) has argued for a recapitulation of the Null Subject Parameter so as to take into consideration diachronic processes of erosion of bound person-number inflection and

[42] See Costa (2011) for evidence that European Portuguese also has topic-comment properties.

[43] *Barlovento* varieties of CVC have the second person forms *bosé* (Veiga 1995:176-177,332-335, 2002:83).

grammaticalization of independent tonic pronouns into SCs. She claims that speakers, in an attempt to economize the processing burdens associated with tracking referents across discourse, are experiencing a constant push-and-pull between forces that seek to maximize the semantic transparency and discourse-pragmatic salience of subject referents, while also minimizing their morphological representation (Figure 3). These forces cause tonic subject pronominals to weaken into atonic SCs, then affixal-person markers, and eventually bound agreement. Bound inflection can also erode to Ø, at which point new tonic forms may be introduced into the system, often by recruitment of existing forms like deictics (Heine & Song 2011).

Figure 3. "The Null Subject Cycle" (Wratil 2011:102)



The "Null Subject Cycle" thus accounts for a wide typological range of partial Null Subject Languages at different diachronic stages of erosion on bound inflectional affixes, grammaticalization on existing pronominals, and reintroduction or recruitment of existing referring expression for use as new tonic subject pronouns. I will return to discussion of the Null Subject Parameter and partial Null Subject Languages in Section 4.5.1., and to the issue of grammaticalization of pronominals in Section 4.5.3.

While it would be useful to examine the quantitative distribution of SCs in CVC using classic syntactic tests of clitichood as was done in Culbertson (2010), I have pointed out that - despite repeated application of the tests for clitichood in various languages over years - there has been little consensus achieved as to the morphosyntactic status and underlying position of SCs in the clause. This is case for CVC, for which Baptista (1995, 2002), following DeGraff (1993) for

Kreyòl Ayisen, has adopted the inflectional affix hypothesis for SCs, whereas Pratas (2004) and Déprez (1994), working with the same clitichood tests, arrived upon the phonological clitic hypothesis for these same languages, respectively.

While many of the studies applying clitichood tests were not quantitative in nature, I will nonetheless forego application of these tests in favor a different approach: I will use inferential statistical analysis consistent with a variationist sociolinguistic methodology to compare the realization SCs in CVC with other anaphoric expressions that also compete in Subject Expression, namely X2SBJs and zero/null subjects. By observing the probabilistic distribution of competing anaphoric elements in the CVC subject domain, I may be able to identify new indirect evidence for the status of SCs. For the time being, I turn to a brief summary of Cardinaletti's & Starke's (1994, 1996, 1999) Typology of Structural Deficiency, since this study was the basis for the tripartite analysis of CVC subject pronominal and makes important claims for the semantic-referential properties of SCs and other subject pronominal elements.

4.3 THE TYPOLOGY OF STRUCTURAL DEFICIENCY (CARDINALETTI & STARKE 1994, 1996, 1999).

In a series of comparative survey studies of Indo-European languages, Cardinaletti & Starke (1994, 1996, 1999) proposed three classes of universally underlying pronominal[44]: strong, weak, and clitic. We have seen that these categories were the basis for the tripartite classification of CVC subject pronouns in Baptsita (2002) and Pratas (2004). The weak and SC classes are characterized by "referential deficiency", that is, their distribution is "impoverished with respect to" strong forms (Cardinaletti & Starke 1994:46; see also Kayne 1975). This referential deficiency is associated with a set of morphosyntactic, semantic, and prosodic properties (45).

[44] While pronouns are the primary focus of their analysis (1994, 1999), they extend their typological categories to other parts of speech such as adverbs and adjectives (1996, 1999), among others. They also emphasize that not all languages necessarily lexicalized all three underlying classes.

(44) "Asymmetries between *strong* and *deficient* pronominals" (Cardinaletti & Starke 1994:46-52)

a.) "*Strong* but not *deficient* forms occur in base or Θ-position (theta position).

b.) *Deficient* forms cannot occur in peripheral position, as in subject-clefts, left or right dislocations, or in isolation.

c.) *Deficient* forms cannot undergo c-modification from adverbs that 'modify the entire noun phrase', nor can they be coordinated with another DP.

d.) *Deficient* pronominals cannot bear contrastive stress, nor occur with ostension (a pointing gesture), unless their referent is already prominent in the discourse, i.e. *deficient* forms get their reference from a prominent discourse antecedent.

e.) Only *deficient* pronominals can occur with expletives or impersonal constructions.

f.) Only *deficient* forms can bear [-human] reference."

After introducing the asymmetries in the distribution of strong and deficient pronominals, Cardinaletti & Starke expand on some of the semantic distinctions related to (44d-e). For example, the observation in (44e) falls-out from the notion that strong pronouns are fully referential, while deficient forms (weak and SC) receive their referential value from a discourse antecedent (Cardinaletti & Starke 1994:52). All forms can bear generic reference, while only deficient pronouns can be used for impersonal or expletive constructions since they are inherently less referential.

After Cardinaletti & Starke established that the morphosyntactic, semantic, and prosodic asymmetries among strong and deficient forms are a consequence of the "referential deficiency"

of deficient forms, they distinguish an additional subset class among the deficient forms, clitic pronouns (1994:60). Clitics are assumed to be even more referentially deficient than weak pronouns, thus establishing a proper subset: weak pronominals are deficient relative to strong forms, and clitics are deficient relative to weak forms. The tripartite typology of pronominals can be represented as in (45).

(45) Hierarchy of pronominal deficiency (Cardinaletti & Starke 1994:60)

More deficient
$\longrightarrow$
Strong > Weak > Clitic

As a result of clitics' greater referential deficiency relative to weak pronouns, they are hypothesized to demonstrate their own set of asymmetries (46) (Cardinaletti & Starke 1994:61-65). The properties listed in (46) were also the basis for testing the clitichood of CVC SCs in Baptsita (2002:245-250) and Pratas (2004:50-58).

(46) "Asymmetries between deficient (weak) and clitic forms" (Cardinaletti & Starke 1994:61-65)

a.) "*Clitics* head an X°-chain, whereas *deficient* forms are a maximal projection (XP).

b.) In instances of doubling, it is *clitics* that are doubled (not weak pronouns).

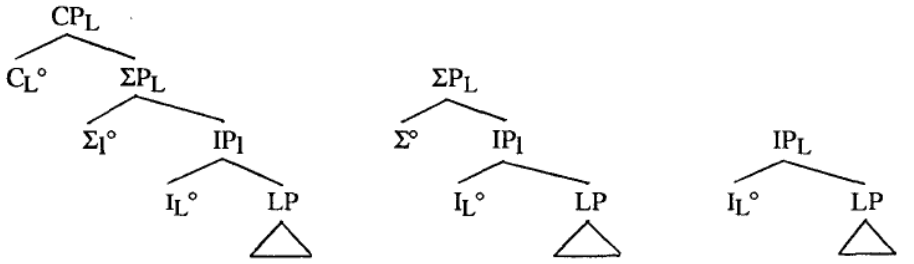
c.) *Clitics* can 'cluster'

d.) *Clitics* have no lexical word-stress, whereas *deficient* pronominals may receive phrasal or contrastive stress"

For Cardinaletti & Starke, referential deficiency results from a literal lack of morphemic structure: strong forms have more functional projections than weak forms, and clitics project less structure than weak pronouns (47a-c). The asymmetric distributional, morphological, semantic, and prosodic asymmetries among strong and deficient forms in (45), and weak and clitic forms in (46), are thus also attributable to these differences in formal underlying structure (1999:177-179;195).

(47) The underlying structure of strong, weak, and clitic pronouns (Cardinaletti & Starke 1994:104; 1999:195)

a. Strong Pronouns b. Weak Pronouns c. Clitic Pronouns



Cardinaletti & Starke (1999:180-185) assume the additional functional projection in strong pronouns comes from an extra morpheme which is lexicalized as an overt 'dummy marker' in some, but not all languages. This 'dummy marker' endows strong pronouns with their referential properties: its presence can be correlated with the properties in (45) like word-stress, the ability to be modified or coordinated, and crucially, [+human] reference (Cardinaletti & Starke 1999:180-182). Interestingly, in some Romance varieties like Italian and Spanish, Cardinaletti & Starke assume this 'dummy marker' morpheme is the preposition *a* ('*a personal*' in Spanish) responsible for the [+/-human] asymmetries in the examples (48a-d) from "Southern and Central Italian dialects" (1999:181).

(48) a.) ***A*** *quella bambina piccola, la metto in primo banco*

DUMMY MARKER this girl small 3.SG.**AN** put.1.SG in first row

b.) ****A*** *quella tavola rossa, la metto vicino alla*

DUMMY MARKER this table red 3.SG.**INAN** put.1.SG near to-

finestra

the window

c.) *Ho parlato* ***a*** *loro.*

Have.1SG spoken DUMMY MARKER 3.PL.**AN**

d.) **Ho aggiunto i pezzi di ricambio* ***a*** *loro*

Have.1SG added the parts of change DUMMY MARKER 3.PL.IN**AN**

Baptista (2002:254) assumes that the initial *a-* on disyllabic CVC strong pronouns (e.g. ***a****mi, **a**bo, **a**el*) is the very same dummy marker; it endows weak pronouns (*mi, bo, el*) with the additional functional projections from (45) responsible for the referential properties of strong forms. Another observation from the Typology of Structural Deficiency that will be important for the present study is that the morpheme underlying null subjects - *pro* - is assumed to be a referentially deficient pronoun (Cardinaletti & Starke 1994:68,89-91).

To summarize, in this subsection I reviewed the Typology of Structural Deficiency (Cardinaletti & Starke 1994, 1996, 1999). This model proposed three universally underlying classes of pronoun: strong, weak, and clitic. Each class is associated the morphosyntactic, semantic, and prosodic properties in (45, above) and (46, above). These properties are assumed to be a direct consequence of the degree of referential deficiency associated with each form; clitics are the most referentially deficient, followed by weak pronouns, and weak pronouns are more referentially deficient than strong forms, which are fully referentially specified (47, above). Because of the referential deficiency of weak and clitic pronominals, they receive their referential specification from coreferential antecedent (or a dummy marker like *a-*, thus becoming strong

forms). Most importantly for the present study, weak and clitic forms are assumed to allow for inanimate reference, as well as other referentially underspecified semantic properties like indefiniteness (or genericity). Since null subjects are taken to be weak pronouns, they should also be found to bear coreference with referentially deficient antecedents. In the next subsection, I turn to study of atonic 'person markers' in 'Probabilistic Grammar' approaches (cf. Claes 2017).

4.4 'PERSON MARKERS' AND ANAPHORIC VS. GRAMMATICAL AGREEMENT.

In 'Probabilistic Grammar' approaches (cf. Claes 2017), like usage-based, functionalist, 'cognitive linguistic', and typology-oriented paradigms, the term 'person marker' is often used for both bound affixes and clitics (Siewierska 2004:24). This overlapping terminology reflects the importance that these approaches attribute to the diachronic source of many bound person affixes cross-linguistically: independent pronouns. Independent pronouns have repeatedly been observed to experience erosion in their syntactic independence and morphophonological form as a result of grammaticalization (Givón 1976; Hopper & Traugott 1993:7; Siewierska 2004:22,263; *inter alia*). Grammaticalization is the diachronic process by which lexical content words come to be progressively more functional and less 'contentful' over time, eventually developing the properties of grammatical markers as a results of processes like semantic bleaching (loss of lexical value) and morphophonological reduction (reduction in the integrity of morphological boundaries, phonological content or prosodic prominence), among others (Givón 1971c; Hopper & Traugott 1993; Bybee, Perkins, & Pagliuca 1994; Campbell & Harris 1995; *inter alia*).

The grammaticalization cline or evolutionary path along which independent pronouns develop into bound inflectional morphology is depicted by Givón (1976) as in (49a-c); a smaller 'slice' of this cline for "dependent person markers" is rendered by Siewierska (2004:22) as in (50). The former representation shows an extra initial step in which non-dependent demonstrative pronouns yield independent tonic pronouns (49a), while the latter representation has an extra final step in which bound inflectional affixes erode into zero-marking.

(49) a.) demonstrative pronoun → stressed independent pronouns

b.) stressed independent pronouns → unstressed clitic pronouns

c.) unstressed clitic pronouns → obligatory verb agreement

(50) weak > clitic > bound > zero

It was mentioned prior (Subsection 4.3) that in Givón's (1976) "NP Detachment Theory", increased reliance on topicalized or topic-shifted constructions with a left-dislocated lexical DP, as in (51), is considered a catalyst for the degradation of independent pronouns towards bound affixes, resulting in the frequent use of X2SBJ constructions.

(51) *Amilson, E ta kumi txeu dimás*

A. […] 3.SG.CL TMA eat a lot too much

'A., he eats way too much.'

As "topic-shifted" constructions like (51) come to occur more frequently in a language, atonic anaphoric subject pronouns/'phonological' SCs may become increasingly grammaticalized; "such topic-shifted constructions become reanalyzed as neutral clauses; the left- […] detached topic becomes the subject and the anaphoric person marker becomes attached to the verb." (Siewierska 2004:263) (52)[45].

[45] The lack of a comma indicates no pause in intonation between the lexical DP and the SC, but instead an unbroken intonational contour. The lack of clitic attachment in (51) represents the not yet fully grammaticalized status of the atonic subject element.

(52) *Amilson E=ta kumi txeu dimás*

A. 3.SG.CL=TMA eat a lot too much

'A. he-eats way too much.'

Elements at the extremes of this grammaticalization cline are seemingly easy to identify and their morphosyntactic status is less controversial. For example, in many varieties of Spanish subject pronouns are clearly stressed independent pronouns while Spanish person-number inflectional is unambiguously 'bound' or 'obligatory verb agreement'.

In many languages, however, the differences between an independent unstressed clitic pronoun and bound inflectional affixes can be blurred (hence the use of the more ambiguous term 'person marker'). One such example is from the Austronesian language Gumawana (Olson 1992:326,306 cited in Siewierska 2004:122); in (53a) the person marker *i* might be said to be a grammatical agreement marker (~bound inflectional affix) given the presence of the lexical DP, whereas in (53b), *i* appears to be an anaphoric person marker (~independent atonic pronoun).

(53) a.) *Kalitoni **i**=pasewa*

K. **3.SG**=work

b.) ***I**=situ vada sinae-na*

3.SG=enter house inside-3SG(INAL)

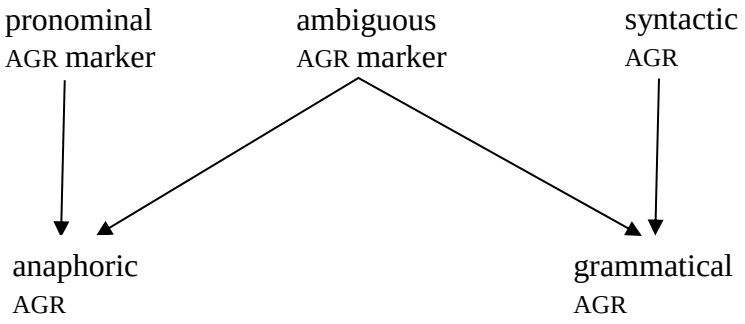
Bresnan & Mchombo (1987) distinguished two different 'types' or 'domains' of agreement in which person markers could participate, these are 'anaphoric agreement' or 'grammatical agreement' (see also Lehmann 1982:219; Siewierska 2004:122). In each domain different 'controllers' regulate agreement on the person marker. When person markers belong to the same syntactic constituent as their 'controller', they are said to be in a local agreement configuration and are subject to 'grammatical agreement' (e.g. bound affixes engaging in person-

number agreement like obligatory Spanish person-number inflection). When person markers are in non-local agreement configurations where they do not form part of the same syntactic constituent as their controller, they are subject to long distance 'anaphoric agreement' (e.g. independent pronominal SCs) (Siewierska 2004:122-127).

As Siewierska (2004:126) observes, person agreement markers like Gumawana *i* seem to be capable of engaging in both types of agreement. In (53a) the person marker is in a local agreement configuration with its controller, the lexical DP *Kalitoni*, while in (53b) the same person marker is in a non-local configuration with its controller (the controller is the antecedent in the prior clause) and is engaged in anaphoric agreement. Thus, person markers like Gumawana *i* seem to lead a 'double life' as pronominal-like anaphors or inflectional-like affixes. Recall from Section 3.3.1 that this is essentially the same conclusion that Zribi-Hertz & Diagne (2002) arrived upon for Wolof person markers.

For Siewierska (2004:126) person markers like Gumawana *i* are 'ambiguous agreement markers'. They can engage in anaphoric (non-local) agreement as well as grammatical (local agreement). Cross-linguistically, ambiguous agreement markers are distinguished from (unambiguous) pronominal agreement markers and syntactic agreement markers. This makes for a tripartite typological distinction in the types of agreement marker available and a bipartite distinction in the kinds of agreement in which person markers can engage (Figure 4).

Figure 4. "Relationship between types of agreement marker and type of agreement." (Siewierska 2004:126)



At middle point of the pronoun-to-affix grammaticalization cline in (49) and (50), morphemes like SCs will often become ambiguous agreement markers (Siewierska 2004:262). SCs might have lost some of their independence as pronominal agreement markers but may have not yet fully transitioned into the status of a syntactic agreement markers bound to the verb. However, languages do not treat the differences between the various kinds of agreement markers categorically: "the distinction between pronominal and ambiguous agreement markers and thus between anaphoric and grammatical agreement is a scalar one." (Siewierska 2004:126). Person agreement is scalar because in some languages, the properties of the controller can determine when and where person agreement markers occur, and those properties usually align with 'accessibility hierarchies' of the kind proposed by Givón (1983b, 2001, 2017), and Ariel (1990, 2001). Siewierska (2004:149) provides a set of "decomposed" sub-hierarchies that reflect the notions assumed in such accessibility hierarchies (54a-e).

(54) a.) the person hierarchy

1ST > 2ND > 3RD

b.) the nominal hierarchy

pronoun > noun

c.) the animacy hierarchy

human > animate > inanimate > abstract

d.) the referential hierarchy

definite > indefinite specific > nonspecific

e.) the focus hierarchy

not in focus > in focus

Siewierska (2004:149) describes this hierarchy of properties of the controller and its relationship to person agreement markers as follows:

> All the hierarchies define a preference for person agreement when the controller exhibits the characteristics on the left of > as compared to those on the right of >. Thus the expectation is that if person agreement is not obligatory in a language, it will occur with controllers displaying the characteristics on the left-hand side of the hierarchies rather than with controllers manifesting the characteristic on the right-hand side.

Person markers at a middle stage of the pronoun-to-affix grammaticalization cline may therefore be variably realized according to the morphosyntactic composition, referential semantic, and discursive function of the controller.

One last important consideration for ambiguous person markers, which as we have seen often have the morphological status of unstressed anaphoric pronouns or SCs, is their functional purpose in discourse organization. The purpose of referring expressions is to introduce or resume discourse referents so that they are identifiable to the interlocutor; different referring expressions index an array of referential properties, such as those related to a referent's semantic and morphosyntactic composition (topicality and salience), or those related to other cognitive domains such as working memory and attention (Givón 2017:27-64). Thus, for Givón (1983a; 2017:5-8), referring expressions are "referent coding devices" that are selected according to the degree to which they signal "referential continuity" (55) (see also Ariel 1990:73-74).

Figure 5. "Referent coding devices and referential continuity" (Givón 2017:6)

highest continuity

a.) zero anaphora

b.) unstressed anaphoric pronouns

c.) stressed independent pronouns

d.) definite NPs

e.) indefinite NPs

f.) modified NPs

lowest continuity

Unstressed anaphoric pronouns "signal maximal referential continuity", meaning they are usually disfavored relative to stressed independent pronouns in contexts of discourse reorientation or introduction of new discourse referents (Givón 2017:7-9), when the antecedent is far away in

numbers of words or clauses (anaphoric distance) (Givón 2017:12-18), or when the antecedent's referent is more salient because it bears morphosyntactic compositional or semantic-referential properties to the left of > in (54) (i.e. it is not 'referentially deficient') (Ariel 1990:64-68; Siewierska 2004:175).

In this subsection, we have seen that in Probabilistic Grammar approaches (cf. Claes 2017), subject elements like SCs are morphemes at an intermediate stage of a grammaticalization cline; at one end lie fully independent pronominals, and the other end are bound inflection affixes (49, above), (50, above). At this middle stage of a longer diachronic change in their formal and functional status, SCs may become what Siewierska (2004) has identified as ambiguous person agreement markers. Ambiguous person agreement markers can engage in local grammatical agreement or non-local anaphoric agreement (Bresnan & Mchombo 1987). With respect to discourse organization, SCs are high referential continuity devices used with highly active, salient, and anaphorically recent discourse referents (Givón 1983a,b,c,d; 2001[1984]; 2017; Ariel 1990, 2001; Siewierska 2004). I turn now to issues of null subjects/zero anaphora, which as we will see are related to the status of SCs and strong/tonic pronominals that we have been exploring so far.

4.5 NULL SUBJECTS/ZERO ANAPHORA IN CVC AND RELATED LANGUAGES.

The second half of this chapter is dedicated to the topic of null subjects, also sometimes called zero anaphora. These terms refer to cases where an anaphoric relationship is express by a phonetically vacuous realization of the structural position usually associated with an overt pronominal anaphor. Research into null subject/zero anaphora (∅ henceforth) has diverged along theoretical lines, with investigations in the GG tradition making different theoretical assumptions than non-GG approaches. These differences in *a priori* assumptions have also led scholars to pose different research questions and draw different conclusions regarding the typology of ∅ anaphora.

The first subsection in this half of the chapter is dedicated the primary line of inquiry that has explored ∅ subjects within the GG tradition, the Null Subject Parameter. Particular attention

is reserved for a group of languages that came to be known as partial Null Subject Languages, with a special focus on Iberian-origin partial Null Subject Languages. I then turn to prior research on ∅ subjects in CVC, which was also conducted through the perspective of the Null Subject Parameter. This is followed by discussion of ∅ subjects in other Iberian-lexifier (and one English-based) creole languages.

Following the discussion of ∅ subjects in partial-NSLs and CVC, I then turn to 'Probabilistic Grammar' approaches; as conceived of in Claes (2017), the Probabilistic Grammar approach combines the theoretical and explanatory frameworks of usage-based, functionalist, 'cognitive linguistic', and typology-driven perspectives, with the quantitative rigor of variationist sociolinguistics and corpus-based research. In this discussion, I will consider studies on Subject Pronoun Expression in Portuguese and Spanish, but also in languages like English and creole languages with little to no inflectional verb morphology that had traditionally been considered non-Null Subject Languages under the Null Subject Parameter.

We will see that, under Probabilistic Grammar approaches to ∅ subjects, many of the same constraints have consistently been found to be active cross-linguistically, and that languages differ mostly in the contexts in which ∅/overt options are allowed to vary, in the relative strength of the effect of independent predictive constraints, and in the ranking of levels within those constraints. I then conclude this chapter by taking stock of the discussion of prior research into SCs, ∅ subjects, and X2SBJs conducted over its course.

4.5.1 The Null Subject Parameter and partial Null Subject Languages in Ibero-Romance.

The original instantiation of the Null Subject Parameter proposed that consistent Null Subjects Languages present a cluster of properties (Perlmutter 1971; Chomsky & Lasnik 1977; Chomsky 1981; Rizzi 1982). Some of these properties include: (i) the availability to express a phonetically unrealized subject that is definite and referential (55a-b); (ii) the availability of free

subject-verb inversion (56a-b); and, (iii) an overt complementizer in *that*-trace constructions (57a-b), among others (Camacho 2008, 2013, 2016; Roberts & Holmberg 2010; *inter alia*).

(55) a.) *Parla italiano* (Italian, +NSP)
b.) **Speaks Italian* (English, -NSP)
'She speaks Italian.'

(56) a.) Hanno telefonato molti studenti (Italian,+NSP)
b.) **Ont téléphoné beaucoup d'étudiants* (French, -NSP)
'Many students have telephoned.'

(57) a.) *Chi hai detto che – ha scritto questo libro?* (Italian, +NSP)
b.) **Who did you say that – wrote this book?*(English, -NSP)
'Who did you say wrote this book?'

Some additional properties that were not initially postulated to cluster with the Null Subject Parameter (NSP henceforth), but that were later found to be related to the NSP, include: (i) null expletives (58a-b); (ii) the inability for overt pronouns to bear arbitrary reference (59a-b); and (iii) adherence to Montalbetti's (1984) *Overt Pronoun Constraint* (60) (Suñer 1983; Jaeggli 1986; Camacho 2008, 2016; *inter alia).*

(58) a. *Llueve* (Spanish, +NSP)
b. *It rains* (English, -NSP)

(59) a. *Dijeron que habían venido* (Spanish, specific or arbitrary reading)

b. *Ellos dijeron que habían venido* (Spanish, specific reading)

'They said that they had come.'

(60) *Todo estudiante$_i$ cree que pro$_i$ es inteligente.*

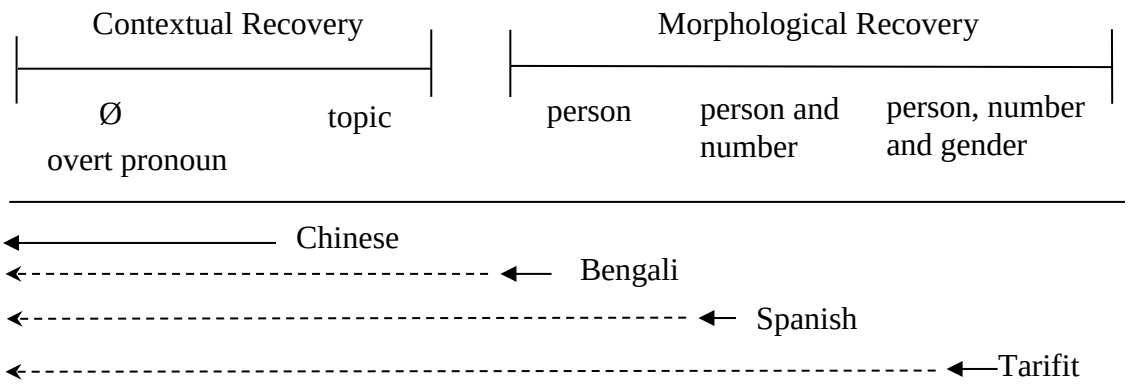
b. *Todo estudiante$_i$ cree que él*$_{i/j}$ es inteligente*

'Every student thinks that he is intelligent.'

It has long been assumed that robust inflectional morphology for subject-verb agreement is the mechanism by which identification of ∅ subjects is carried out in consistent Null Subject Languages (NSLs henceforth). However, this has not been found to apply without exception cross-linguistically, and challenges in formalizing the concept of "rich inflectional morphology" have prevented this notion from being accepted as an inherent feature of NSLs (Camacho 2013:111). For some NSLs, like topic-comment or discourse-oriented languages, ∅ subjects are identified by discourse topics (cf. Huang 1984; *inter alia*).

Cole (2009) and Camacho (2013:111-114) have shown that even in consistent NSLs, when ∅ subjects are not identified by inflectional morphology, they can alternatively be identified by a discourse topic. Cole (2009) and Camacho (2013:111-114) assume that languages have different "minimal morphological thresholds" on the identification of Ø subjects. Identification proceeds by a "cascading strategy": "[…] null subjects are identified if uniquely identified by agreement morphology. If that fails, they are identified by reference to an antecedent context, and if that is not possible, overt pronouns are used." (Camacho 2013:112-113). Languages lie on a gradient scale that places different limitations of the threshold necessary for Ø subject identification (61).

(61) "Recoverability scale" for the identification of Ø subjects (Cole 2009; Camacho 2013:113)



Several Iberian-origin vernaculars in the Americas such as Brazilian Portuguese (BP), Dominican Spanish (DS) and Afro-Bolivian Spanish (ABS), among many others, have been identified as partial NSLs. This classification has been asserted due to the lower overall rates of Ø subjects in these languages, the use of overt subjects in seemingly redundant or discursively infelicitous contexts (i.e. non-contrastively), and because these languages adhere to only some of the properties associated with the NSP cluster (cf. Duarte 1993, 1995; Toribio 1993, 2000; Barbosa, Kato, & Duarte 2005; Camacho 2008, 2013, 2016; Holmberg *et al.* 2009; *inter alia*). The use of non-contrastive, non-emphatic overt subject pronouns can be observed for BP (62a) (Barbosa, Duarte, & Kato 2005), DS (62b) (Toribio 2000), and ABS (62c) (Sessarego & Rodríguez-Riccelli 2018).

(62) a.) *E* ***ele$_i$*** *precisou ir ao banheiro. Quando* ***ele$_i$*** *viu o que que era o banheiro,* ***ele$_i$*** *ficou apavorado* (BP)

'And he had to go to the bathroom. When he saw what the bathroom looked like he was terrified.'

b.) *Si* ***ellos$_i$*** *me dicen que* ***yo$_j$*** *estoy en peligro cuando* ***ellos$_i$*** *me entren la aguja por el ombligo* ***yo$_j$*** *me voy a ver en una situación de estrés.* (DS)

'If they tell me that I am in danger when they put the needle in my belly-button, I am going to find myself in a stressful situation.'

c.) ***Yo$_i$*** *no tengo plata,* ***yo$_i$*** *no quiere comprá.* (ABS)

'I do not have [any] money, I do not want to buy [it].'

Each of these varieties has also undergone reduction in their respective inflectional morphological paradigms. In the case of the BP, Duarte (1993, 1995, 2000) demonstrated that overall rates of overt subjects increased dramatically over the course of nearly a century and half, from about 20% overt subjects in the mid-19th century, to 74% in 1992 (Figure 6) (Duarte 1993; Barbosa, Duarte, & Kato 2005). The increase in overt subject pronouns in BP was argued to be motivated by changes in the subject pronoun inventory (Table 27), which in turn triggered a reduction in the inflectional morphological paradigm (Table 28) (Camacho 2008; adapted from Duarte 1993, 1995:15).

Figure 6. Overt subject pronouns in BP mid-19th to late 20th centuries (Barbosa, Duarte, & Kato 2005; adapted from Duarte 1993:112)

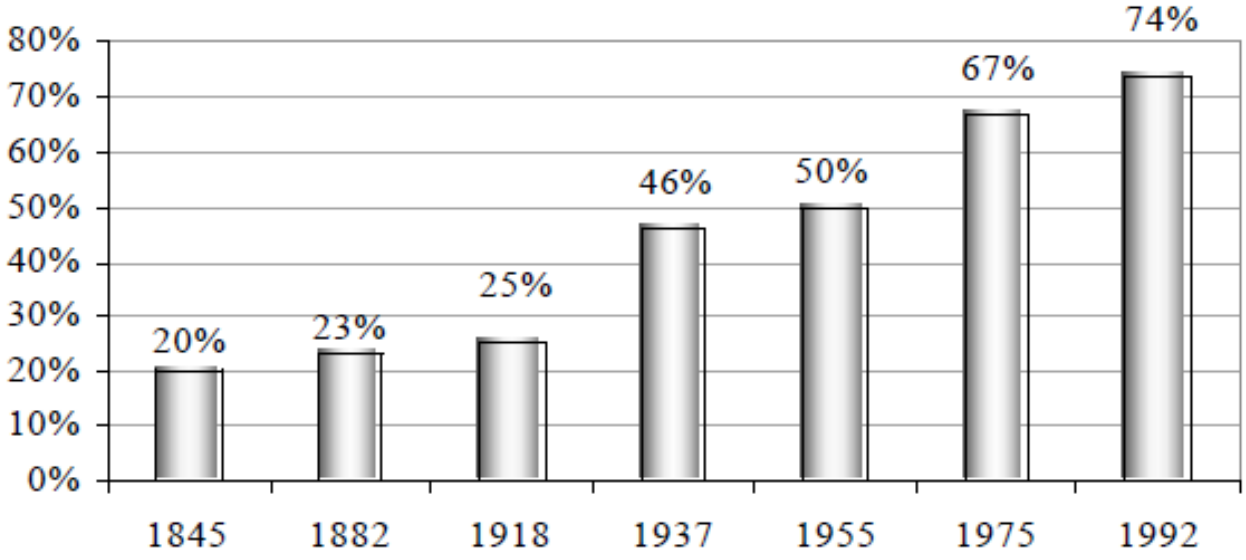


Table 27. The evolution of the verbal inflection paradigm in BP (adapted from Duarte 2000:19 and Camacho 2008, 2016)

Person-number	Paradigm 1	Paradigm 2	Paradigm 3
Speaker.SG	*am-o*	*am-o*	*am-o*
Hearer.SG	*am-a-s*	--	--
Other.SG	*am-a*	*am-a*	*am-a*
Speaker.PL	*am-a-mos*	*am-a-mos*	--
	--	*am-a*	*am-a*
Addressee.PL	*am-a-is*	--	--
	am-a-m	*am-a-m*	*am-a-m*
Other.PL	*am-a-m*	*am-a-m*	*am-a-m*

Table 28. The evolution of the subject pronoun inventory in BP (adapted from Camacho 2016)

Person-number	Pronoun		Verbal ending
Speaker.SG	*eu*	*eu*	*am-o*
Addressee.SG	*tu*	você	*am-a*
	vós		
Other.SG	*ele/ela*	*ele/ela*	*am-a*
Speaker.PL	*nós*	*a gente*	*am-a*
Addressee.PL	*vós*	*vocês*	*am-a-m*
	vocês		
Other.PL	*ele/ela*	*ele/ela*	*am-a-m*

In DS, reductions in the verbal inflection paradigm were triggered by phonological changes that implicate processes of weakening in syllable-final rhotics and sibilants that are associated with person-number inflection morphemes or infinitive markers (62a-d) (Toribio 2000).

(63) a. Norm: [sal.ˈtaɾ], [ˈsal.tas],[sal.ˈta.ßãn]

b. Santo Domingo: [sal.ˈtal], [ˈsal.ta], [sal.ˈta.ßãŋ]

c. Cibao Valley: [saj.ˈtaj], [ˈsaj.ta], [saj.ˈta.ßãŋ]

d. Southern coast: [saɾ.ˈtaɾ], [ˈsaɾ.ta], [saɾ.ˈta.ßãŋ]

In the most basilectal varieties of ABS, the verb is invariably 3rd person singular, though subject-verb agreement in terms of overt inflectional morphology is highly variable and sensitive to effects for register, age of speaker, tense, and person-number[46] (Sessarego 2017:174-175). Reduction of the inflectional paradigm in ABS has been attributed to regular SLA processes, such as morphological bottlenecks (cf. Slabakova 2006, 2008), or syntactic phenomena that lie at the cognitive interfaces between domains of grammar (cf. Sorace 2003, 2011; Tsimpli & Sorace 2006; Rothman & Slabakova 2011; Jackendoff 2015, *inter alia*). When certain population structures obtain in a given language community, these SLA processes can be iterated over successive generations, which can trigger a process of "L1 acquisition (nativization) of advanced L2 grammars" by which the variable features associated with these L2 acquisition are incorporated into the grammar across generations (Sessarego 2011, 2013, 2014a, 2019; Rao & Sessarego, 2016, 2018; Sessarego & Rao 2016; Sessarego & Ferreira 2016; Sessarego & Gutiérrez-Rexach 2017; Sessarego & Rodríguez-Riccelli 2018).

Another change in BP, DS, and ABS associated with the NSP cluster of properties includes the loss of subject-verb inversion in *wh*-questions, a property that BP lost between the 19th and early 20th centuries (63a-b) (Silva 2001; see also Kato & Duarte 2003; Sessarego & Gutiérrez-Rexach 2017). Non-inversion in questions also occurs in DS[47] (64a-b) (Toribio 2000), though rates

46 This is also the case for BP and DS, as will be discussed ahead when addressing variationist sociolinguistic studies on Portuguese and Spanish subject expression.

47 Non-subject-verb inversion, as well as several of the properties discussed for DS, can be found in all Caribbean varieties of Spanish; the productivity of inversion in each varies depending on the variety in question and social context under consideration, among many other factors (cf. Martínez Sanz 2011, and references therein; Camacho 2016:33, and references therein)

of inversion and its acceptance in grammatically judgement tasks is a point of ongoing debate (Toribio 1993; Ordóñez & Olarrea 2006; Camacho 2016; *and references therein*); inversion also obtains variably in ABS (65a-d) (Sessarego & Rodríguez-Riccelli 2018).

(64) a.) *O que **a Maria** leciona?*

‘What does Maria teach?’

O que leciona **a Maria?*

b.) *Onde **ela** leciona?*

‘Where does she teach?’

Onde leciona **ela?*

(65) a.) ¿*Qué número **tú** anotaste*?

‘What number did you write down?’

b.) ¿*Qué **yo** les voy a mandar a esos muchachos*?

‘What am I going to send to those boys?’

(66) a.) ¿*Qué **oté** tá jugá*?

‘What are you playing?’

b.) ¿*De ondi **oté** viene*?

‘Where are you from?’

c.) ¿*Y cuándo vuelve **(oté)***?

‘And when are you coming back?’

d.) ¿*Di ondi son **(otene)***?

‘Where are you from?’

Changes in the interpretation of overt pronouns as per the Overt Pronoun Constraint (Montalbetti 1984) have been reported for BP, in that an overt pre-verbal subject pronoun can be

interpreted as a bound variable (67) (Barbosa, Duarte, & Kato 2005). ABS speakers also accepted overt pre-verbal subject pronouns with a bound interpretation (68) (Sessarego & Rodríguez-Riccelli 2018).

(67) *[Ninguém no Brasil]$_i$ acha que ele$_i$ é prejudicado pelo Governo.*
'Nobody in Brazil thinks that he is harmed by the Government.'

(68) *Ele$_i$ piensa que ele$_{i/k}$ es inteligente.*
'He thinks he is intelligent.'

Modesto (2000:152-153) argued that the interpretation of ∅ subjects in BP differs from that of an overt pronoun in that the former must be locally bound by a c-commanding antecedent see (also Camacho 2008, 2016). In (69a), the ∅ subject (*pro*) can be bound by the subject preceding the verb, but not by the post-verbal object, while in (69b) the interpretation of the embedded overt subject pronoun is ambiguous, as it can be bound by either the preceding subject, object, or another referent in the discourse. In (69c) the ∅ subject can be bound by a *wh*-operator that has undergone fronting, and in (69d) the same holds for a left-dislocated object.

(69) a.) *O Paulo$_i$ convenceu o Pedro$_j$ que pro$_{i/*j/*k}$ tinha que ir embora.*
'Paulo convinced Pedro that he had to go away.'
b.) *O Paulo$_i$ convenceu o Pedro$_j$ que ele$_{i/j/k}$ tinha que ir embora.*
'Paulo convinced Pedro that he had to go away.'
c.) *Quem$_i$ que o Pedro$_j$ convenceu t$_i$ que pro$_{i/?*j}$ tinha que ir embora?*
'Who did Pedro convince that he had to go away?'
d.) *A Maria$_i$, o Pedro$_j$ convenceu t$_i$ que pro$_{i/j/k}$ tinha que ir embora.*
'Pedro convinced Maria that he/she had to go away'

These data led Modesto (2008) to argue that BP has developed some of the properties of topic-drop, topic-prominent, or discourse-oriented languages (cf. Li & Thompson 1976; Huang 1984), in which ∅ subjects can be identified by discourse topics (see also Camacho 2013). Barbosa, Duarte, & Kato (2005), arguing against Modesto (2000), maintain that in BP ∅ subjects can be referential and are available in root clauses (as opposed to only occurring in bound embedded contexts, see also Camacho 2016), and represent a residual usage of ∅ subjects following a diachronic change in BP away from a consistent Null Subject Language.

With respect to the semantic properties associated with each class of pronominals under the Typology of Structural Deficiency, we noted that the referential deficient nature of weak pronouns allows them to resume antecedent that bear inanimate, nonspecific, or indefinite reference (Cardinaletti & Starke 1994, 1996, 1999). Importantly, Cardinaletti & Starke take *pro* to be a deficient pronoun, meaning that ∅ subjects are available to resume semantically referentially deficient antecedents. This is precisely what Kato & Duarte (2003, 2005) and Duarte & Soares da Silva (2016) have found for BP; ∅ subjects are favored with inanimate and nonspecific antecedents. This prompted them to propose an "avoid referentially deficient pronoun" constraint based on Montalbetti's (1984) OPC.

As can be gleaned from BP, DS, and ABS - all varieties which have their diachronic origins in consistent NSLs - restructuring in the subject domain interacted with reductions in the verbal inflection paradigm with respect to person/number morphology[48], and this reduction is thought to have impacted the overall rates of ∅ subjects, as well as the distribution and interpretation of overt and ∅ subjects with respect to the NSP cluster of properties. In this subsection we have seen that the NSP had to be modified to account for partial NSLs, including several Romance vernaculars that experienced diachronic changes in the pronominal and verbal domains to the extent that a partial NSL classification has become warranted. The formal mechanism for identifying ∅ subjects

[48] The loss of distinctive verbal inflectional morphology has also been invoked to explain the shift in Old, to Middle, to Modern French from a consistent NSL to a non-NSL (cf. Adams 1987; Roberts 1993; Hulk & van Kemenade 1995; Kaiser 2009).

was reworked as a language-specific 'minimum morphological threshold' where ∅ can be discourse-identified in lieu of identification from inflection morphology (61) (Cole 2009, Camacho 2008, 2013). This allows for the NSP to account for some cross-linguistic variability in languages' expression of the NSP cluster of properties, and we saw examples of variability with respect to these properties in BP, DS, and ABS. Later in this chapter, however, we will see that the NSP still fails to capture the broader 'variable rule system' of interacting probabilistic constraints underlying variable Subject Pronoun Expression. Beforehand, I turn to a discussion of ∅ subjects in CVC and other Iberian-lexifier creoles.

4.5.2 Null subjects in Cabo-Verdean Creole and other Ibero-Romance-lexifier creoles.

The classification of CVC as an NSL has been disputed primarily as a consequence of the analysis attributed to SCs. Pratas (2004) and Costa & Pratas (2008, 2013) identify CVC SCs as independent subject pronominals in the canonical subject position[49] (SPEC,IP/TP) (i.e. the phonological clitic hypothesis) (see 34b, above).

(70) [IP [DP ***E***][I' [VP *papia*]]]

By extension, in X2SBJ constructions, the lexical DP or strong pronoun is analyzed as left-dislocated (cf. Pratas 2004:96) (71). This perspective is effectively equivalent to the phonological clitic hypothesis in that CVC utterances with a lone SC do not involve any ∅ subject anaphor in the subject position[50] (see 43b above).

(71) [CP ***Jon***$_i$[IP [DP ***E***$_i$][I' [VP *papia*]]]

[49] This analysis would also have to be extended to the Gumawana example in (53a).
[50] This analysis would also have to be extended to the Gumawana example in (53b).

With respect to the NSP, Pratas (2004) and Costa & Pratas (2008, 2013) classify CVC as a partial NSL. Costa & Pratas (2013) deny that ∅ subjects are possible in root referential contexts. They also observe that CVC only adheres to some of the NSP cluster of properties: inversion only obtains with indefinite subjects of unaccusative verbs (71a-c), CVC has ∅ subjects in non-argumental constructions like weather predicates[51] (72) and impersonals (73), and CVC does not exhibit *that*-trace effects (74).

(72) a. *(dja) txiga tres omi.*
TMA arrive three man
'There arrived three men.'

b. **(dja) txiga Djon/*(dja) more Djon.*
TMA arrive Djon TMA die Djon

c. *Djon (dja) txiga./Djon (dja) more.*
Djon TMA arrive Djon TMA die
'Djon arrived.'/'Djon died. '

(73) *Sata txobe na Lisboa.*
TMA rain in Lisbon
' It's raining in Lisbon.'

(74) *Li pode fumadu.*
LOC may smoke.PASS
'One/you may smoke here.'

[51] Note that CVC also has the construction *txuba sata txobi* 'rain is raining'.

(75) *Kenha ki bu ta pensa ma kunpra livru ?*
who COMP 2.SG TMA think COMP buy book
'Who do you think bought the book? '

Based on comparisons of CVC data with Modesto's (2000, 2008) BP data, Costa & Pratas (2013) concluded that the only instances of (argumental) *pro* in CVC are as a bound variable, available only in embedded contexts when the ∅ subject can be bound by a quantifier (76) or a *wh*-operator (77).

(76) a. *Kenha$_i$ ki ta atxa ma Ø$_i$/*$_j$ kore faxi labanta mo.*
who$_i$ COMP TMA think COMP 3.SG$_{i/*j}$ run fast raise hand.
'Whoever thinks he/she ran fast raise (their) hand.'
b. *Kenha$_i$ ki ta atxa ma el*$_{i/j}$ kore faxi labanta mo.*
who$_i$ COMP TMA think COMP 3.SG $_{i*/j}$ run fast raise hand
'Whoever thinks he/she ran fast raise (your) hand.'

(77) a. *Ningen$_i$ ka atxa livru ki Ø$_i$/$_{*j}$ perdeba.*
no.one$_i$ NEG find book that 3.SG$_{i/*j}$ lose.PST
b. *Ningen$_i$ ka atxa livru ki el$_{*i/j}$ perdeba.*
no.one$_i$ NEG find book COMP 3.SG $_{i*/j}$ lose.PST
'No one has found the book that he/she lost.'

From this perspective, both CVC and BP only have 'true' ∅ subjects (*pro*) in embedded contexts, though Modesto (2008) attributes *pro*-drop patterns in BP to an ongoing change in this language towards patterns of subject expression that resembles topic-drop languages. For Costa & Pratas (2008, 2013), topic-drop properties are absent in CVC, leaving [argumental] *pro* to emerge

only in embedded contexts where the subject position is bound by an operator, but never in root contexts or when bound by a referential lexical DP (Costa & Pratas 2013).

As a proponent of the inflectional affix hypothesis, Baptista (1995, 2002) argues that CVC is an NSL. To reiterate, the inflectional affix hypothesis classifies SCs as tense heads in the inflectional layer (AGR/INFL). In other words, SCs are analyzed as agreement affixes rather than 'true' subject pronouns in a canonical subject position (SPEC,IP/TP). In conjunction with analyses adopting the NSP under which AGR is responsible for the identification of ∅ subjects (*pro*), the SC is taken to be the 'spell-out' of AGR and therefore can identify ∅ in the subject position[52] (78) (see 34a above). Following this line of reasoning, Baptista (2002:266) classified CVC as a "radical *pro*-drop language", equivalent in status to consistent NSLs under the NSP.

(78) [IP [DP ***pro***][I' ***E*** [VP *papia*]]]

Consequently, the co-occurrence of both a lexical DP (or tonic subject pronoun) and SC is 'true doubling' in the sense of Culbertson (2010); the lexical DP is in the canonical subject position while the SC remains in the inflection layer[53] (79) (see 34a above).

(79) [IP [DP ***Jon$_i$***][I' ***E$_i$*** [VP *papia*]]]

In addition to taking utterances with a lone SC in pre-verbal position as instances of genuine ∅ subjects, Baptista (2002:259-260) also provides numerous examples from her corpus of CVC utterances where there are no overt subject elements whatsoever occupying the pre-verbal position, leaving the 'subject slot' devoid of phonological content (anaphoric ∅).

52 This analysis would also have to be extended to the Gumawana example in (53a).
53 This analysis would also have to be extended to the Gumawana example in (53b).

She finds that this occurs most often with copular verbs. With individual-level predicate *e*[54] 'to be' the ∅ subject receives a "default" 3rd person interpretation (80a-b), whereas with stage-level predicate *sta* the ∅ subject can be interpreted in any person/number instantiation (81b).

(80) a.) *(El) é nha pai*
(3SG.STR) COP my father
'He is my father

b.) *(El) é spertu*
(3SG.STR) COP expert
'He is clever'

(81) a.) Bu sta livri
2SG.CL COP free
'You are free.'

b.) Ø sta livri
pro COP free
'I/you/he/she/it/we/they is/are free'

Beyond these contexts, she also found ∅ subjects with non-copular verbs (Baptista 2002:258); these occurred most often in the 3rd person, which she attributed to the ability of AGR to check "'abstract' 3rd person features" as a default procedure (82)-(83).

(82) *Kantu ki lion di matu ben góra, Ø txiga, Ø fla…*
when COMP Lion of wood come ADV *pro* arrive *pro* say
'When the Lion from the Woods came then, [he] arrived, [he] said…'

(83) *Nho Lion ka faze almusu, dja Ø ba faze pastoría.*
mister Lion NEG make lunch, TMA *pro* go do field
'Mister Lion did not make lunch, he had gone to the fields'

(84) [AGRSP [NP pro_i][AGR $_i^{3RD}$][VP *txiga*]]

54 Recall that *e* is the only verb in CVC for which SCs are categorically excluded.

Though less productive, she also found examples of Ø subjects where the verb is in other person-number configurations (84)-(85) (Baptsita 2002:259). However, it remains unexplained how Ø subjects in the 2nd and 1st person examples would be formally derived.

(85) *Ami pur akazu, N teni oitenti sinku anu, Ø nasi na mil novsenti kinzi, Ø teni oitenti sinku anu*

'I, incidentally, am eighty-five years old, I was born nineteen fifteen, I am eighty-five years old.'

(86) *N bai nha kaza dja, N bai pega na vivensia pa konta na nha kaza. Panha lenha na montadu, bendi… Ø bendi kel fixinhu di lenha, Ø ba trabadja djenti, Ø ganha kel dinhirinhu, Ø ben kunpra kel kafizinhu.*

'I went to my own house then, and went to seek a livelihood relying on my home. I would take the wood in the grove and would sell it. I would sell that little piece of wood, I would go to work over people's houses, I would earn a small sum, I would buy a little coffee.'

(87) *Bo bu ta konxi biblia, Ø ta konxi biblia sagrada, Ø ta konxe-l*

'You, you know the bible, you the sacred bible, you know it.'

(88) *Mo la é sima Merka, kes arvi$_i$, si txuba sta ku bentu, si arvi rabenta, Ø$_i$ da na bo… la me bu fika.*

'Those trees, if it's raining and windy, if the trees fall, [they] fall on you…'

Costa & Pratas (2013) counter that Ø subjects of the sort found in (80)-(88) are ambiguous between generic/impersonal and genuinely argumental readings; they reject them as genuine instances of 'root' Ø subjects.

Ø subjects similar to (82)-(88) have also been attested in other creole languages, including several Iberian-lexifier creoles and at least one English-based one. Mufwene (1988) observed Ø subjects in the English-lexifier creole Gullah. Like many creole languages, Gullah has Ø subjects in expletive constructions and with arbitrary and impersonal referents, all contexts which he identifies as instances of "big PRO" and "$pro_{\text{ARB(ITRARY)}}$". He also presents data showing that referential Ø subjects are possible provided the adequate discursive context in which they can be recovered (identified), usually when bearing coreference with preceding antecedent (89a-e) (Mufwene 1988:238-239).

(89) a.) [ʌ kɛ̃ driŋk də waʲn // *pro* giv mi hɛdɛʲk]
'I can't drink the wine ... [it] gives me a headache'

b.) [ʌ ɛ̃ si: him *pro* mʌs bi ovə tu yiəz]
'I haven't seen him [for it] must be over two years'

c.) [aʲ no lɛʲdi dɛd // *pro* go tu lɛʲdi fyunrəl]
'I know Lady is dead ... [I] went to Lady's funeral'

d.) [*pro* θiŋk ʃi ho:m naʷ]
'[I] think she [is] home now'

e.) [wi də ho: // ho: ɛn pik kɔtn // ɛnihaʷ *pro* də du ɛvritiŋ in fi:l]
'We used to hoe ... hoe and pick cotton… Anyhow, we used to do everything in the field'

Reflecting on the validity of the NSP, Mufwene (1988) compares Ø subject use in Gullah with consistent NSLs like Spanish and Italian, discourse-oriented or topic-drop NSLs like Mandarin Chinese, and with non-NSLs like English. He concludes that the NSP is too rigid to describe the gradient patterns of Ø subject use across these languages: Ø subjects in Gullah are far more restricted than in consistent NSLs since Gullah lacks the rich inflectional morphology assumed to allow for identification of Ø subjects. At the same time, Gullah is also more restricted

than discourse-oriented NSLs like Mandarin since the discursive contexts permitting ∅ subjects in Gullah are more limited. Yet still, ∅ subjects in Gullah are freer than in English, where ∅ subjects often occur in tag-questions, and in 'diary-style' or written registers (cf. Haegeman 1990).

Another example of a creole language permitting ∅ subjects despite lacking rich person-number agreement inflection comes from a member of the Upper Guinea Creole Portuguese group, Papiamentu. Kouwenberg (1990) considers the interpretational differences among overt and ∅ subjects in subordinate complement clauses following complementizer *pa*, which is used with non-factive predicates in Papiamentu (90). As we have seen in Table 20 and (31)-(33) (Section 4.1, above), CVC has the same complementizer *pa* which shares the semantic and selectional properties of the homophonous Papiamentu complementizer. Kouwenberg concludes that in these contexts, Papiamentu has available referential *pro*, and in a broader range of contexts (such as root clauses), non-argumental covert categories PRO and pro_{ARB} underlie ∅.

(90) a.) *Cheli$_i$ a pidi Feli$_j$ pa e$_{j/k}$ bai playa kuné*
'Estel has asked Felix if he/she would go into town with her.'
b.) *Cheli$_i$ a pidi Feli$_j$ pa Ø$_j$ bai playa kuné*
'Estel has asked Felix to go to town with her.'

Lipski (1999) considers several Iberian-lexifier Creoles, including Chabacano, Palenquero, and Papiamentu. In his data, there can be observed limited use of ∅ subjects in each of these languages. For each, he posits that ∅ subjects are usually 'null constants' like PRO, or pro_{ARB}, but that some referential ∅ subjects are possible.

In the case of Chabacano, he finds that identification of ∅ subjects involves a combination of strategies from the superstrate and substrate languages. Like in its superstrate (Spanish), ∅ subjects in Chabacano can get an impersonal 3rd person reading. However, substrate influence allows for discourse-identified referential ∅ subjects (91a-c), while all other instances of ∅ subjects can be interpreted as expletive, impersonal, arbitrary, or generic (92a-c).

(91) a.) *Mama talya na bentana Ø ta espera konmigo* (McKaughan 1954:215)
'Mama was in the window [and she was] waiting for me'

b.) *Ya lyama el rey kon el baw ya Ø pregunta konele porké ele ta karga su kasa* (McKaughan 1954: 216)
'The king called the turtle [and he] asked him why he carried his house'

c.) *Kosa el ya ase ya sake su korta-pluma ya Ø empesa pone aguhero na buli del mana olya* (McKaughan 1954: 222)
'What did he do, [he] took out his penknife [and he] began putting holes in the bottom of the pots'

(92) a.) *pro$_{ARB}$ ta siña kanila "English"*
'[they/people] teach them English'

b. *pro$_{ARB}$ necesita pa gat syempre usa chabacano*
'[one/you/people/they] still need to use Chabacano all the time'

c. *pro$_{ARB}$ ta manda kortá kon ese palay, pro$_{ARB}$ ta asé kamaring grande, pro$_{ARB}$ alyá ta juntá palay ...*
'[you/one] have/has the rice cut, [you/one] make[s] big piles, [you/one] gather[s] the rice up there'

Neither Palenquero nor Papiamentu have a substrate with discourse-oriented ∅ subjects (Lipski 1999). For this reason, Lipski claims, discourse-identified ∅ subjects are not possible in these languages; seemingly-referential ∅ subjects, he argues, are actually interpretationally ambiguous between with impersonal/arbitrary/expletive ∅, or are cases of coordination or serial verb constructions (93a-d) (Lipski 1999).

(93) a.) *Suto á sé limpiá yuka, Ø sé limpiá aló; Ø ejperá kodte aló, Ø é sé kotá aló*

'We clean the yucca [and we] clean the rice; [we] wait for the rice harvest, [and we] cut the rice' (Palenquero)

b.) *é sé polé pasá ri aí nu. Ø sé salí nu.*

'He can't get past here. [he] can't get out.' (Palenquero)

c.) *Tur hende a kore drenta kas i Ø bai lur ya yalusí kiko ta pasa*

'Everybody ran home and [they] went to the window to see what was happening' (Papiamentu) (Maurer 1988:105)

d.) *tin kaminda di kana, maske Ø ta poko smalitu*

'There is a footpath, although [it] is a little narrow' (Papiamentu) (Maurer 1988:134)

In this subsection, we have seen two competing analyses for the status of CVC under the NSP. One analysis (Pratas 2004; Costa & Pratas 2008, 2013) takes CVC to be a partial NSL since SCs are assumed to be independent pronouns in the canonical subject position and since CVC does not adhere to all of the properties assumed to cluster under the NSP. This approach rejects that Ø subjects are possible in root contexts in CVC, concluding that they are limited to highly specialized embedded contexts where they can be bound by a c-commanding indefinite or WH-operator.

The competing analysis (Baptista 1995, 2002) adopts the inflectional affix hypothesis by categorizing SCs in CVC as tense heads, situated in the inflectional layer, thus implying that the subject position contains Ø in clauses with a singleton SC. Further, Baptista (2002) offers numerous examples of Ø subjects where there are no overt subject elements whatsoever present. In these examples (80)-(88), Ø subjects either receive a default interpretation based on copula selection or are coreferential with some prior discourse antecedent.

These examples bear a strong resemblance to Ø subjects observed in other creole languages. I corroborate Baptista's observations of Ø subjects with additional examples from my

own corpus. In the meantime, I turn to issues of ∅ subjects and Subject Pronoun Expression or anaphora resolution in Romance, but also cross-linguistically, particularly in languages lacking 'rich' inflectional morphology for identifying ∅ subjects like English.

4.5.3 'Probabilistic Grammar' approaches to Subject Pronoun Expression/anaphora resolution.

Following the proposal in Claes (2017) for a 'Probabilistic Grammar' that merges variationist sociolinguistic approaches with those from the usage-based, functionalist, 'cognitive linguistic', and typology-oriented approaches, in this subsection, I explore issues of Subject Pronoun Expression (SPE; also called anaphora resolution) from the above theoretical paradigms. I begin with research on Spanish, Portuguese, and English SPE in variationist sociolinguistics. Then, I explore theories on ∅ subjects in usage-based, functionalist, and 'cognitive linguistic' approaches that are generally concerned with issues of anaphora resolution as a function of antecedent accessibility, among other cognitive processes associated with tracking referents across a discourse.

Subject Pronoun Expression and ∅ subjects have been among the most enduring topics of investigation in the study of Spanish morphosyntax within the variationist sociolinguistic paradigm. Variationist sociolinguistic approaches, from an epistemological perspective, explore how structured variation is represented in one's knowledge of a language. To do so, variationist approaches adopt a quantitative methodology to identify the set of independent predictor variables (also called independent factors or constraints), both language-internal and external, that probabilistically condition the realization of a particular linguistic variant. The outcome of the linguistic variable under consideration is represented as a dependent variable (sometimes called a response variable or outcome variable) in inferential statistical analyses (cf. Weinreich, Labov, & Herzog 1968; Labov 1972, 2004; Cerdegen & Sankoff 1974; Tagliamonte 2006, 2012, *inter alia*).

Thus, variationist studies on Spanish Subject Pronoun Expression (SPE, henceforth) take the [non-]realization of a subject pronouns like *nosotros*, with a finite referential verb such as *hablamos* in the utterance *(nosotros) hablamos*, to be the dependent variable under analysis, with the realization of ∅/OVERT PRONOUN as possible outcomes. All contexts in which one of these outcomes is possible, and where variation can be observed in congruent contexts, are recorded. Then, a probabilistic model is constructed, and the aggregate of the dependent variable outcomes are regressed on the set of independent predictor variables chosen for investigation. The independent predictor variables represent factors that have been postulated or previously observed in the literature to have some conditioning influence on the realization of the dependent variable.

Over the years, the use of regression to conduct a variable rule analysis has been enhanced by the use of multivariate analysis, which in addition to providing information about the overall statistical significance of an independent factor group, the relative ranking of the levels within that group, and the strength of the effect of a factor group relative to the other factor groups considered, can also help to identify interactions among independent variables, and can accommodate both discrete and continuous independent factors (Tagliamonte 2006, 2012; Martínez Sanz 2011:176-178). With the advent of mixed-effects modeling, it is also possible to account for random effects, which mitigate the residual variance in the model, or the kind of variance associated with individual participants or data collection materials that cannot be account for within the fixed-effects' variance (cf. Johnson 2008; Tagliamonte 2012; Tagliamonte & Baayen 2012).

Studies on Spanish SPE have been conducted on a wide range of dialectal varieties, but have also employed different methodological innovations and sometimes consider a different set of independent variables for analysis, which has at times made comparability and replicability of studies a formidable challenge (cf. Cameron 1993; Travis 2007; Prada Pérez 2009; Martínez Sanz 2011). Despite these challenges, studies on Spanish SPE have consistently been at the avant-garde of the field of variationist sociolinguistics.

One major contribution to the field from this research program has been the realization that while overall rates of [un-]expressed subject pronouns vary greatly cross-dialectally, all varieties

of Spanish, not just those presumed to be in flux or influenced by other languages, experience variable SPE (Martínez Sanz 2011:175). What is more, a consistent set of independent variables appears to condition SPE across varieties of Spanish, regardless of the overall rate of [non-]occurrence of ∅ subjects in a given variety (Martínez Sanz 2011:175; Carvalho, Orozco & Lapidus Shin 2015:xiv-xv). Instead, varieties of Spanish tend to differ in the relative strength of the effect from each independent variable group. Thus, the locus of variation and cross-dialectal differences is not usually found in the overall set of independent variables implicated in conditioning SPE, but rather in the magnitude of influence that each constraint asserts relative to the others, the interactions that might obtain among the predictors, and the ranking of levels within those predictive constraints (Otheguy, Zentella & Livert 2007; Martínez Sanz 2011; Otheguy & Zentella 2012; Carvalho, Orozco & Lapidus Shin eds. 2015).

Though a comprehensive summary of the various factors found to influence SPE in Spanish across varieties and social contexts is beyond on the scope of this dissertation, it will be useful to identify some of the major factors involved; some of these will be revisited and described in more detail in the next, methodological chapter (see Martínez Sanz 2011:174-180; Carvalho, Orozco, & Lapidus Shin 2015:xiv-xv; for comprehensive summaries). Among the generalizations regarding the influence of independent variables on SPE cross-dialectally in Spanish, social factors such as age, gender and occupation do not appear to exert much influence, or at best an inconsistent influence, though in situations of language contact differences for social factors do emerge (Martínez Sanz 2011:178; and references therein).

Among the language-internal factors found to consistently achieve significance in quantitative studies on Spanish SPE, both in monolingual and language contact settings, are: grammatical PERSON/NUMBER, SWITCH-REFERENCE/COREFERENCE, TOPIC CONTINUITY, DISTANCE FROM PREVIOUS MENTION of the referent in the discourse (anaphoric distance), PRIMING by the surface morphological form of the antecedent, the TMA MORPHOLOGY of the verb associated with the subject in question, the LEXICAL SEMANTICS of the verb, and the CLAUSE TYPE in which the subject is found (Martínez Sanz 2011:174-180, and references therein; Otheguy & Zentella 2012

and references therein; Carvalho, Orozco & Lapidus Shin 2015:xiv-xv, and references therein; Claes 2017).

Quantitative analyses of SPE in Portuguese have mostly focused on the interpretive properties of ∅ subjects and their identification conditions in BP, rather than on conducting variable rule analyses *per se*. We have seen that Duarte (1993, 1995, 2000) and Barbosa, Kato & Duarte (2005), using corpus data, identified an ongoing process of change in BP whereby diachronic alterations in the subject pronoun inventory and in verbal morphological inflection paradigm triggered major reductions in the overall rates of ∅ subjects over the course the late 19th and 20th centuries (and a loss of the NSP cluster of properties). Comparison between SPE in BP and European Portuguese also reveal different patterns for the person-number of the referent; for instance, in BP the highest rate of overt subject pronouns is in the 2nd person while in European Portuguese it is in the 1st person (Duarte 2000).

Studies of SPE in African varieties of Portuguese are few and do not easily lend themselves to comparison due to methodological differences among them and some contradictory results that likely stem from those methodological differences. For instance, Oliveira & Ferreira dos Santos (2007) found that, despite increased use of 2nd person singular *você*[55] and 1st person plural *a gente*, which like 3rd person Portuguese verbs are not marked with distinctive inflection, Angolan Portuguese (AP) still retains 2nd person *tu,* which does receive the distinctive inflectional suffix *-s*. Perhaps consequentially, Oliveira & Ferreira dos Santos find overall rates of ∅ subjects that are comparable to European Portuguese (EP henceforth). Like EP, AP has the most overt subject pronouns in the 1st person. Teixeira (2012) found contradictory results with rates of overt pronouns in AP comparable to those found for BP, but as Bouchard (2018:121-122) points out, the methodologies applied in these studies are under-described, rendering comparability and reproducibility impossible. Dias (2009) examined the written output of bilingual Portuguese-Changana speaking 5th grade students for rates of SPE in Mozambique. She found a 52.5% overall

[55] It is unclear if Angola Portuguese *você* retains the politeness value associated with it in EP, or if it is becoming employed in familiar usage like in BP.

rate of Ø subjects, with high rates of overt subject pronouns in the 1st person singular, but high rates of Ø subjects in the 1st plural.

Bouchard (2018) conducted a variationist sociolinguistic analysis of SPE in Santomean Portuguese (StP henceforth). She collected and analyzed data from 48 sociolinguistic and ethnographic interviews from Portuguese speakers in São Tomé, the capital city of São Tomé & Príncipe. Her application of a mixed-effects logistic regression returned as significant a collection of independent factors similar to those that have been found relevant for conditioning SPE in Spanish, namely: CLAUSE TYPE, PRIMING, MORPHOLOGICAL REGULARITY, SEMANTIC CONTENT, PERSON/NUMBER, ANIMACY, and COREFERENTIALITY.

Of the language-external factors she analyzed, EDUCATION LEVEL and the random factor SPEAKER achieved significance, though the factors AGE, GENDER, and ETHNIC ORIGIN did not. She found overall rates of overt SPE approaching those of EP at 68.5%. The general insignificance of social factors in conditioning SPE in StP reveals a trend similar to variationist studies on variable SPE in Spanish. Bouchard attributes the relevance of EDUCATION LEVEL to the use of the formal register among university-educated participants, who because of an association between propositional form and surface syntax, are thought to attribute prestige to speech with high rates of overt subject pronouns. Based on the close comparability of the SPE variable rule system in StP with those of EP and Spanish, Bouchard concludes that despite effectuating changes in other domains, StP remains conservative with respect to patterns of SPE (2018).

In a recent study that blends the terminology and theoretical constructs of the Null Subject Parameter with a variationist sociolinguistic methodology, Duarte & Soares da Silva (2016) compare several corpus-based datasets for SPE in European and American varieties of Romance. The authors probed for differences across varieties in the independent factors found to condition Ø/overt variation in Italian, European Spanish (ES), EP, Argentine Spanish (AS), Puerto Rican Spanish (PRS), and BP. They argue that there are gradient differences in the extent to which these varieties can be classified as NSLs, according to the factors returned as significant from a series

of variable-rule analyses. As a broad generalization, they argue that, with respect to verbal inflection paradigms:

> (a) the fewer oppositions a paradigm exhibits, the higher the tendency for a system to exhibit overt subjects and (b) the change toward overt subjects, caused by a reduction in the number of oppositions, cannot be explained as functional compensation in the system analyzed" (Duarte & Soares da Silva 2016:1).

The assertion in (a) has been conceived of before by Roberts (1993), for whom a "functionally rich" verbal inflection paradigm can contain up to two syncretism for there to be identification of ∅ subjects (Duarte & Soares da Silva 2016:2). Assertion (b) refers to the notion that overt subjects are used to compensate for ambiguous inflection, such as when overt subjects are used more in the imperfect and subjunctive in Iberian NSLs, a claim that is controversial both in generativist paradigms (Camacho 2016) and variationist sociolinguistic ones (Torres Cacoullos & Travis 2017, 2019).

In their variable-factor-based categorization, they describe Italian as the most 'prototypical' or consistent NSL, followed by ES and EP, which though also best classified as NSLs, exhibit a less stringent adherence to the properties associated with the NSP when compared to Italian. After ES and EP comes AS, which itself approximates less a consistent NSL when compared with ES and EP. In turn, PRS approximates a consistent NSL even less so than AS. Finally, BP is the language that least resemble a consistent NSL (it is the best example of a partial NSL among the languages under investigation). This renders the hierarchy Italian > ES > EP > AS > PRS > BP, with Italian on the left representing the most consistent NSL and each language to the right incrementally representing languages in various stages of change away from an NSL towards a partial NSL.

For Italian, Duarte & Soares da Silva (2016:5-9) make use of the dataset in Marins (2009). This analysis consisted of conversational speech collected from Italian speaking participants stratified for age and education between the ages of 18 and 59. All contexts with contrastive and emphatic subjects were removed, as these are thought to trigger an overt subject in consistent NSLs (Rizzi 1989). After the removal of these contexts, the data exhibited a rate of 14% overt subjects,

and all inanimate referents were categorically null; following Marins (2009), Duarte & Soares da Silva claim that the residual 14% overt subjects are best explained by discursive factors such as the topic shift or direct reference to another participant in the conversation (2016:6).

The variable-rule analysis in Marins (2009) returned PERSON and ANTECEDENT SYNTACTIC ACCESSIBILITY as the most significant factors conditioning Subject Pronoun Expression. The factor ANTECEDENT SYNTACTIC ACCESSIBILITY (Calabrese 1986) was determined by assigning one of five patterns to each subject under consideration for analysis (Marins 2009; following Soares da Silva 2006; Barbosa, Duarte, & Kato 2005). Each pattern refers to the clausal configuration and anaphoric relationship between the subject under analysis and its antecedent. It also encodes the syntactic role of the antecedent and clausal adjacency. These patterns will be explored in the next chapter since they partially inform one of the constraints considered in the present study.

ES and EP are close to Italian in bearing the properties of an NSL, but differ from Italian in that they are sensitive to a different set of structural patterns for the constraint ANTECEDENT SYNTACTIC ACCESSIBILITY (Duarte & Soares da Silva 2016:9-14). In ES, the overall rate of ∅ subjects was 76%, and like Italian, the 3rd person was the most favorable context for ∅ subjects (88%, FW = 0.71), followed by 2nd person informal *tú* (78%, FW = 0.41), polite 2nd person *usted* (69%, FW = 0.23), and 1st person *yo* (65%, FW = 0.38) (Duarte & Soares da Silva 2016:9-14).

For ANTECEDENT SYNTACTIC ACCESSIBILITY, in Italian, the distance between the subject under analysis and its antecedent was the sole relevant factor in determining the optimal structural pattern that most favored ∅ subjects; in ES and EP, a c-commanding relationship also becomes a relevant factor (Duarte & Soares da Silva 2016:10). EP patterns similarly to ES, with an overall ∅ subject rate of 72%, though for EP, GRAMMATICAL PERSON was not found to be a relevant factor conditioning SPE. Still, the relative ranking of ∅ subjects by frequency from most to least is the same as for Italian and ES: 3rd person > 2nd person > 1st person (Duarte & Soares da Silva 2016:11).

Interestingly, the authors find a rate of 6% overt pronouns with an inanimate referent. Therefore, like CVC and BP, EP also employs the same set of overt subject pronouns for animate or inanimate discourse referents. As for the structural patterns in the constraint ANTECEDENT

SYNTACTIC ACCESSIBILITY, EP shows a pattern similar to those of ES, though the presence of the antecedent within the same sentence is of greater importance in EP than in ES. Thus, ES and EP pattern together in overall rate of ∅ subjects, and in the importance of a c-commanding relationship between an antecedent and the coreferential subject. EP differs from ES and Italian in permitting low rates of overt subjects with inanimate referents (Duarte & Soares da Silva 2016:12-13).

Moving to the American varieties of Spanish and Portuguese, each one was found to exhibit lower overall rates of ∅ subjects than European Romance varieties. AS and PRS retain ∅ subjects in more than half of the observations (68% and 59%, respectively), while BP has rates of ∅ subjects far below the rest (29%). Whereas AS and PRS never use overt pronouns for inanimate referents, in BP (like in CVC and EP), overt pronouns are available to resume antecedent bearing inanimate or non-specific reference (Duarte & Soares da Silva 2016:14-19). For AS, ∅ subject rates are over 50% in all persons, being highest for the 3rd person (81%, FW = 0.69), and second lowest for the 1st person (63%, FW = 0.46), followed by polite 2nd person *usted* (60%, FW = 0.35). In PRS, the range in rates of ∅ subjects per grammatical person, and the FWs associated with each, are lower than for AS and the other varieties in question, suggesting a diminishment in the importance of grammatical person in conditioning SPE in PRS.

AS and PRS also differed in their patterns of ∅ subject realization according to the structural patterns for antecedent accessibly (Duarte & Soares da Silva 2016:15-16). AS patterns with ES and EP with respect to ANTECEDENT SYNTACTIC ACCESSIBILITY: the distance between a subject and its antecedent is relevant, though less so than in Italian, while a c-commanding antecedent is also relevant. In PRS, a c-commanding antecedent is the only condition still favoring the use of ∅ subjects. Thus, Duarte & Soares da Silva (2016:15-16) consider the c-commanding context a "resistance context" in that it is the only syntactic reflex preventing PRS from 'fully' becoming a partial NSL (Duarte & Soares da Silva 2016:15-16).

BP is the variety that patterns least like the others; it is the furthest from a consistent NSL and is at an advance stage of an ongoing diachronic change towards a partial NSL. In the data presented in Duarte (1993, 1995), in no grammatical person configuration does BP exceed a rate

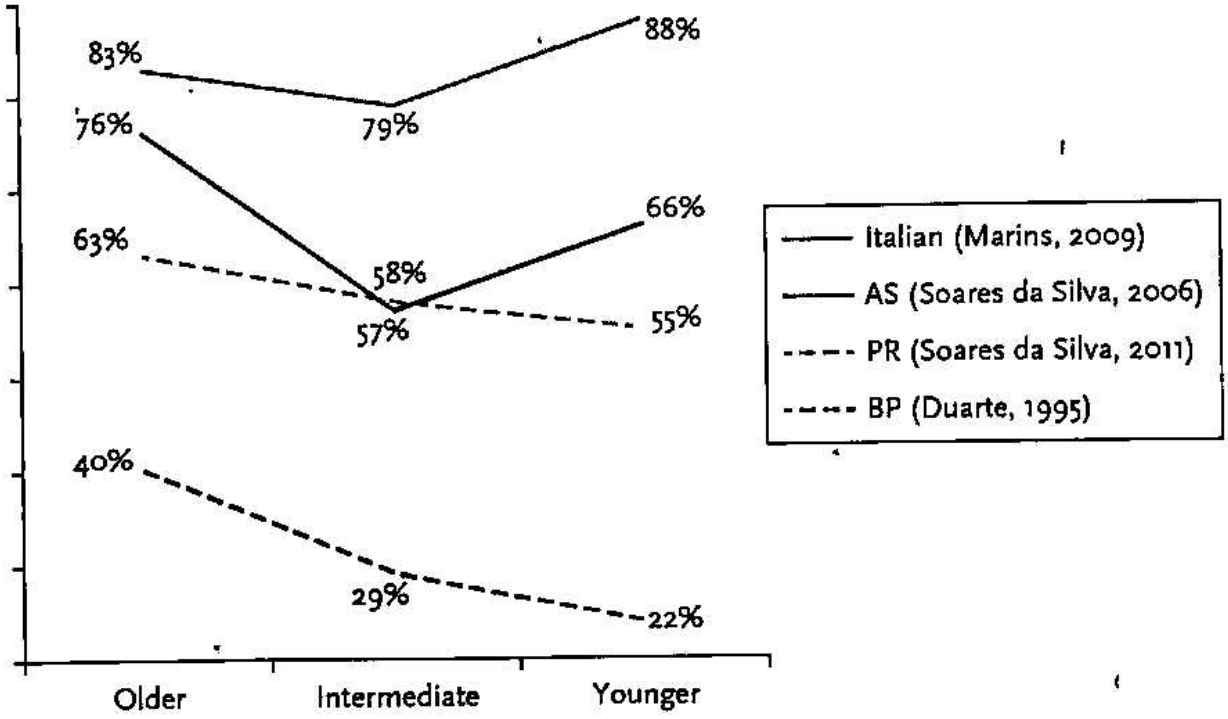
of 50% ∅ subjects. The highest rates of ∅ subjects was in the 3rd person (41%, RW = 0.39), the lowest was in the 2nd person (10%; FW = 0.61), and the 1st person was between these (26%, RW = 0.61) (Duarte & Soares da Silva 2016:15-16). The authors also find that most 1st person singular ∅ subjects occurred in matrix rather than embedded contexts.

Since in BP the same overt subject pronouns may refer to [+/-animate] and [+/-specific] referents, the authors coded 3rd person subjects for each possible combination. They found that, with non-specific reference, ∅ subject rates were equal to or exceeded 50%. ∅ subjects occurred at a rate of 56% with [+human/-specific] referents (RW = 0.66), and a rate of 50% with [-human/-specific] referents (RW = 0.70). [-human/+specific] referents exhibited ∅ subjects at rate of 41% (RW = 0.60) and ∅ subject rates were lowest with [+human/+specific] referents at 33% (RW = 0.40) (Duarte & Soares da Silva 2016:17-18). Thus, the more referentially specified the antecedent, the greater the probability that the coreferential subject will be overt.

Recall that Kato & Duarte (2003, 2005) proposed an "avoid referentially deficient pronoun" constraint based on similar results for the semantic properties of discourse antecedents (Section 4.5.1, above). As we saw in the subsection on X2SBJs (4.2, above), BP also allows X2SBJ constructions when the lexical DP involved bears inanimate and non-specific reference (cf. Tavares Silva, Carvalho & Ziober 2018). Of 52 tokens of X2SBJs in Duarte's (1995) data, 21 (=40%) bore [-human/+specific] reference, 4 (=7%) bore [+human/-specific] reference, and 3 (=5%) bore [-human/-specific] reference (2016:19).

The clearest evidence indicating that BP and PRS are in a state of transition from an NSL to a partial NSL comes from the authors' data on age, which was another factor selected as significant in their multivariate analysis. An apparent time analysis shows V-curves for Italian and ARS, indicating they are in a relatively stable state with respect to their status as NSLs, while BP and PRS show a steadily decreasing rate of ∅ subjects from older to younger age groups (Figure 7).

Figure 7. Effects for age in patterns of Ø/overt Subject Expression in several Romance varieties (Duarte & Soares da Silva 2016: 20).



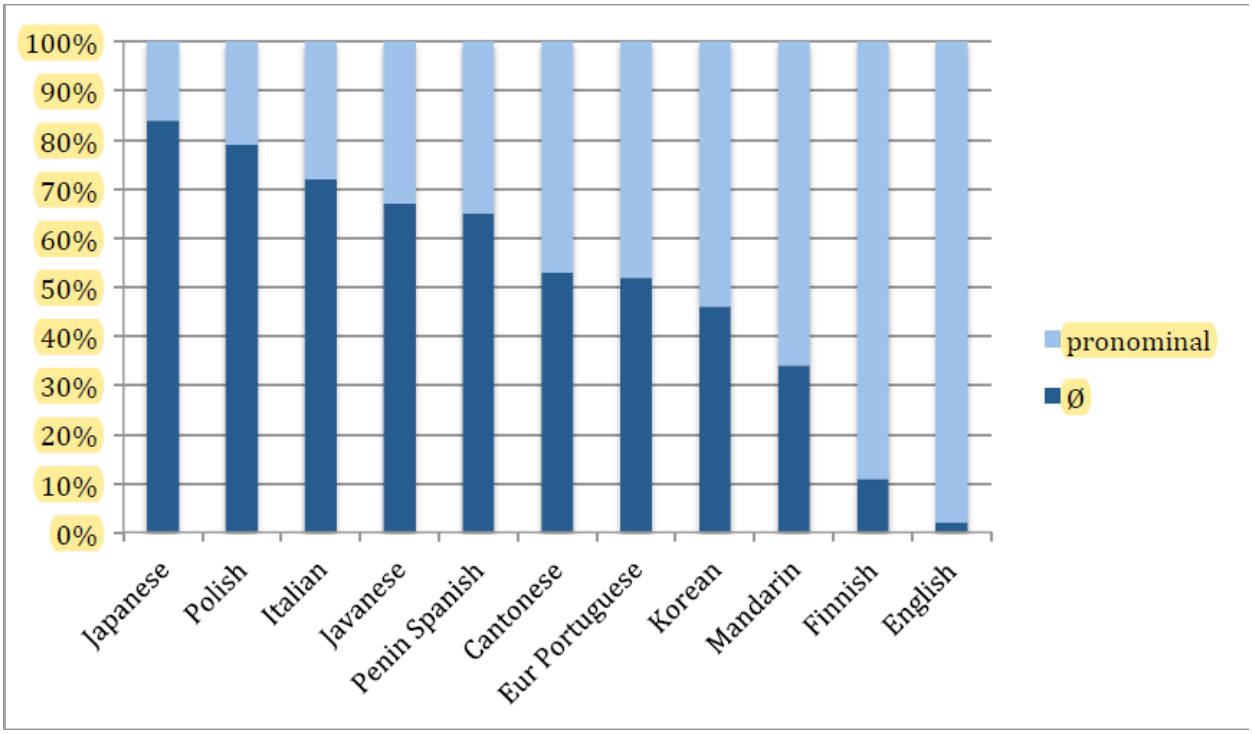
This new perspective on the NSP shows that the cross-linguistic properties of NSLs cannot be fully understood without considering the probabilistic effects of predictor factors that condition variable SPE, particularly in the case of partial NSLs. Duarte & Soares da Silva (2016:22-23) take the change in progress in BP to result from the "embedding of (micro)parametric changes", which following Kato (1999), Kato & Duarte (2013, 2014), and Duarte & Kato (2014), involves the combination of two general process affecting the status of overt pronominals in BP: "weakening inflectional morphology allows weak, clitic-like pronouns adjoined to the head of TP and leaves its specifier (Spec TP) free to receive topic-like subjects".

This is essentially equivalent to the 'NP Detachment Hypothesis' proposed by Givón (1976), where a historical increase in the reliance of topic-shifted left-dislocation of lexical DPs triggers grammaticalization on independent subject pronominals, driving them towards clitic and eventually affixal status. One proposal was mentioned that attributed SC status to reduced forms of overt subject pronouns in BP (Luís & Kaiser 2016) (Section 4.4, above); I also examined Wratil's

(2011) 'Null Subject Cycle' model (see Figure 4, Section 4.4, above) that captures the diachronic interplay between the subject and verbal domains as bound person-number inflectional morphology erodes, tonic pronominals grammaticalize into SCs, and SCs increasingly take-on affix-like status.

As we have observed thus far, in variationist sociolinguistic studies on SPE in Portuguese and Spanish - regardless of the overall rates of Ø subjects across varieties of these languages and irrespective their status under the NSP - the same set of predictive constraints emerge as significant in conditioning the variable rule systems that underlie patterns of SPE. These constraints include PRIMING, the morphological composition of the antecedent (PERSON/NUMBER, SURFACE FORM OF THE ANTECEDENT), CLAUSE TYPE, and the syntactic role of the antecedent in its containing clause combined with the clausal configuration between anaphor and antecedent (ANTECEDENT ACCESSIBILITY, ANAPHORIC DISTANCE). Without accounting for the effects of these constraints, overall rates of Ø subjects are superfluous; consider the cross-linguistic rates of SPE for 1st person singular referents compiled from various independent studies in Torres Cacoullos & Travis (2019; and references therein) (Figure 8).

Figure 8. “Rates of 1sg subject expression (Ø vs. pronominal) across different languages” (from Torres Cacoullos & Travis 2019)



These facts led Torres Cacoullos & Travis (2019:3) to assert that: “Rates of use are thus equivocal, being susceptible to the preponderance or dearth in a data set of some propitious context, which may be fortuitous or due to extra-grammatical, situational considerations, such as register or topic.”

Furthermore, even in languages that have traditionally been thought to be non-NSLs like English, ∅ subjects are nonetheless attested; for instance, Haegeman (1990) observed the use of ∅ subjects in ‘diary’ and colloquial registers of English[56] (2, repeated here for convince in 94 below).

(94) *A very sensible day yesterday. Ø saw no one. Ø took the bus to Southwark Bridge. Ø walked along Thames Street.*

[56] See also (Givón 1976, 2001[1984], 2017; Harvie 1998; Haegeman & Ishane 1999; Bailey 2011; Torres Cacoullos & Travis 2014, 2015, 2016, 2019; Wagner 2016; Travis & Lindstrom 2017; *inter alia*)

Despite the low rates of ∅ subjects and impoverished person inflection in English, it has long been observed that the same set of predictive constraints that are active in variable SPE in Portuguese and Spanish, also condition the distribution of ∅ subjects in English (once the variable context has been delimited) (cf. Givón 1976, 1983d, 2001[1984], 2017; Ariel 1990; Harvie 1998; Bailey 2011; Torres Cacoullos & Travis 2014, 2015, 2019; Wagner 2016; Travis & Lindstrom 2017; *inter alia*). Under approaches like those of Givón (1992b; 2001[1984]; 2017), Chafe (1987, 1994), and Ariel (1990, 2001), among others, grammatical cues are understood to be "mental processing instructions" that allow for the interlocutor to track and process referents across a discourse. There are several hierarchical levels of 'discourse coherence' for structuring and organizing discourse (e.g. story, episode, paragraph), but the smallest and most basic unit is the 'clause-chain' (Givón 2017:5) or the 'Intonational Unit' (Du Bois *et al.* 1993; Chafe 1994:53-70; Torres Cacoullos & Travis 2019). An Intonational Unit (IU henceforth) is defined as "segments of speech uttered in a single, coherent intonation contour." (Du Bois *et al*. 1993:47; Chafe 1994:58-60; Torres Cacoullos & Travis 2019).

In transcribed speech, each IU occupies a line of a transcript and ends with a punctuation mark; a period indicates final intonation characterized by a drop to low pitch, and a comma indicates continuing intonation defined by a rising, sustained, or slightly falling pitch. Continuing intonation at the end of IU serves to string or link IUs (or clause chains) together into larger 'discourse chunks' (~paragraphs). The prosodic linking of IUs is also associated with greater referential continuity (Givón 2017:5-6) and increased activation of discourse referents (Chafe 1994:71-74), both conditions that are associated with a greater likelihood of employing a ∅ subject over an overt one (cf. Givón 1976, 2001[1984], 2017; Ariel 1990:58-61; Torres Cacoullos & Travis 2014, 2015, 2016, 2018, 2019; Wagner 2016; Travis & Lindstrom 2017). Similarly, when English IUs are syntactically linked with a coordinating conjunction (VP coordination), Torres Cacoullos & Travis (2014, 2015, 2019) have demonstrated that SPE is variable. In particular, syntactically linked IUs promote ∅ subjects, especially when linked with the conjunction *and* (Torres Cacoullos & Travis 2016:6-11).

The IU transcription procedure allowed Torres Cacoullos & Travis to examine the effects of LINKING across coreferential subject IUs on variable SPE under four conditions: (i) maximal continuity, when the antecedent's IU is syntactically and prosodically linked to anaphor's IU (95a-b); (ii) intermediate continuity, when IUs with coreferential subjects are only syntactically linked (96a-b); (iii) intermediate continuity, when IUs with coreferential subjects are only prosodically linked (97a-b); and, (iv) minimal continuity, when IUs with coreferential subjects are unlinked (98a-b) (2019:9).

(95) ✓ syntactic linking (*and*)
✓ prosodic linking (same Intonation Unit)
*I wrote home to my family and **Ø said**,*
*... and then I go and **I talk** to him.*

(96) ✓ syntactic linking (*and*)
✓ prosodic linking (continuing (comma) intonation contour)
*... Dad called him**,** and **Ø told** him he had to.*
*so he came**,***
*and **he stood** opposite me,*

(97) ✓ syntactic linking (*and*)
✗ prosodic linking (e.g., final intonation contour)
a.) *and he ran them off**.***
*... and **Ø saved** their lives.*
b.) *he's a broker**.***
*... And **he buys** hay*

(98) ✗ syntactic linking (no *and*)

✓ prosodic linking (continuing intonation contour [=comma])

a. *And then I worked for a rancher over there for a while**,***

*... **Ø followed** the rodeos for a while,*

b. *That's what I did all day today**,***

***I had** ... three or four different kids come up,*

(99) ✗ syntactic linking (no *and*)

✗ prosodic linking (e.g., final intonation contour [=period])

a. *... And yesterday was the first day she used it**.***

***Ø Put** a bunch of stuff in it to read,*

b. *... I do the hard labor**.***

***I build** barns and,*

By attending to IUs, Torres Cacoullos & Travis were able to identify the variable context or envelope of variation for variable SPE in English: with the exception of IUs syntactically linked by *and*, the English ∅ subject is restricted to an IU-initial position (2019:10-12). This finding is supported by evidence from a variationist study on colloquial English 1st person subjects that found a first-subject turn-initial favoring effect for the realization of anaphoric ∅ (Wagner 2016:24).

Thus, the variable context for English SPE was identified as consisting of (i) coreferential subjects linked by either syntactic or prosodic linking; and, (ii) subjects in IU-initial position in main declarative clauses. Any non-coordinated non-IU-initial positions were excluded, as well as contracted forms, discourse markers, and non-coordinated quotatives (Torres Cacoullos & Travis 2019:14). This delimitation of the envelope then allowed Torres Cacoullos & Travis to compare patterns of SPE between English and Spanish. The Spanish envelope of variation for SPE is of course much broader, Torres Cacoullos & Travis excluded only wh-questions and emphatic

constructions of the type subject pronoun + *sí* + verb (2019:14). Thus, the envelope of variation for SPE is a locus of cross-linguistic variation; languages will differ in the syntactic, discursive, and prosodic contexts in which SPE is free to vary (Torres Cacoullos & Travis 2019:14).

After establishing the variable contexts for datasets in both languages, the authors examined rates of SPE in each (within each language's respective envelope of variation) (Torres Cacoullos & Travis 2019:14). They compared two independent logistic regressions, one for the English dataset and the other for the Spanish dataset; the correspondences between the resulting variable rule systems for both languages are remarkable. They found the strongest effects for the factor linking; adjacent IUs containing a subject and its antecedent that were both syntactically and prosodically linked (96, above) were the most favorable condition for ∅ subjects in both languages. This was followed by priming; a ∅ antecedent was found to trigger the expression of a ∅ subject in both languages. Finally, for the constraint verb class, dynamic verbs were found to favor the expression of ∅ subjects in English and in Spanish (Torres Cacoullos & Travis 2019:14-17).

The results for SPE in English from Torres Cacoullos & Travis (2014, 2015, 2019) have largely been corroborated by other studies on SPE in colloquial English such as Wagner (2016). Wagner also found priming ('persistence') to be the most important constraint with strongest favoring effect on 1st person ∅ subjects in English. Her findings also support those of Torres Cacoullos & Travis for verb class, and the previously mentioned first-subject turn-initial promoting effect on the realization of a ∅ subject. In addition to these factors, Wagner also found the constraint verb phrase complexity[57] to promote ∅ subjects.

Corpus and variationist sociolinguistic studies on SPE in English like Torres Cacoullos & Travis (2014, 2015, 2016, 2018, 2019) and Wagner (2016), in Portuguese and Spanish (cf. Otheguy, Zentella & Livert 2007; Martínez Sanz 2011; Otheguy & Zentella 2012; Carvalho, Orozco & Lapidus Shin eds. 2015; Duarte & Soares da Silva 2016; Claes 2017; Bouchard 2018;

[57] More complex VPs have more elements like modals, auxiliaries, tense-mood-aspect encoding, or combinations of these (Wagner 2016:14-16).

inter alia), and other languages (cf. Speyer 2016 for Old High German), together have been revealing that the same set of probabilistic constraints are active in conditioning SPE cross-linguistically. This is the case regardless of the overall rates of ∅ subjects in a given language, and irrespective of a language's adherence to the NSP cluster properties. Where languages differ is in the delimitation of variable context for SPE (Torres Cacoullos & Travis 2019), in the strength of the predictive constraints, interactions among predictors, and in the ranking of levels within those constraints (cf. Carvalho, Orozco & Lapidus Shin eds. 2015). These findings lend robust quantitative support to what those in the usage-based, functionalist, typological, and 'cognitive linguistic' traditions have observed and theorized all along: constraints like priming, anaphoric distance, discourse organization, and the morphosyntactic status of subject antecedents, all conspire to regulate antecedent accessibility as part of the way we process discourse in speech, and thus how anaphora are selected (cf. Givón 1983a, 2001[1984], 2017; Ariel 1990; Chafe 1994).

One final aspect of antecedent accessibility that has been of great theoretical importance in usage-based, functionalist, and 'cognitive linguistic' approaches to the study of ∅ subjects, is the semantics of discourse referents. We have already explored the notion that referring expressions are "referential coding devices" that signal different degrees of referential continuity (cf. Givón 2017:6) (Section 4.4, above). Recalling (55, above, repeated here for convenience 100), the two referring expressions with the highest referential continuity are scs and ∅ subjects.

(100) "Referent coding devices and referential continuity" (Givón 2017:6)

highest continuity

a.) zero anaphora

b.) unstressed anaphoric pronouns

c.) stressed independent pronouns

d.) definite NPs

e.) indefinite NPs

f.) modified NPs

lowest continuity

Thus, referring expressions like SCs and ∅ subjects are used when their antecedents are most activated and salient (and therefore most accessible). The conditions for activation/salience involve many of the constraints that we have been examining in this subsection: the syntactic role and morphological composition of the antecedent (PRIMING, ANTECEDENT ACCESSIBILITY PATTERN, PERSON/NUMBER, SURFACE MORPHOLOGICAL FORM OF THE ANTECEDENT) and the clausal and coreferential relationship between the anaphor and its antecedent (ANAPHORIC DISTANCE, ANTECEDENT ACCESSIBILITY PATTERN, CLAUSE TYPE, SWITCH REFERENCE).

Cross-linguistically, ∅ subjects should be at their most probable when they are maximally accessible (and thus maximally salient/activated): when they are primed by a coreferential ∅ subject antecedent in an immediately prior (or within the same) IU/clause-chain, such that the subject and its antecedent are in prosodically and syntactically linked IUs. While SCs are also probable under these conditions, they are usually promoted over ∅ subjects in contexts of switch-reference and at clause-chain boundaries (in the former context, tonic forms are usually promoted over atonic ones, but this depends on a confluence of probabilistic conditions related to antecedent discontinuity) (Givón 2017:7).

To this set of conditions, theoreticians like (Givón 1983a:10; 2017:6-7) and Ariel (1990) add the semantic properties of the antecedent. Generally speaking, indefinite, inanimate, and nonspecific referents should be less salient and thus less accessible, presumably disfavoring ∅ subjects. This classification can be observed in accessibility hierarchies like (54, Section 4.4, above, repeated here for convenience 101), in which indefinite NPs are associated with low referential continuity, as in the 'animacy' hierarchy (100c) and 'referential' (100d) hierarchy (Siewierska 2004:149,174-200).

(101) a.) the person hierarchy

1ST > 2ND > 3RD

b.) the nominal hierarchy

pronoun > noun

c.) the animacy hierarchy

human > animate > inanimate > abstract

d.) the referential hierarchy

definite > indefinite specific > nonspecific

e.) the focus hierarchy

not in focus > in focus

In this subsection, we have seen that Probabilistic Grammar approaches (Claes 2017) - which encompass usage-based, functionalist, 'cognitive linguistic', typological, variationist sociolinguistic, and corpus-driven approaches to morphosyntax and discourse organization - have converged on a set of overlapping constraints that appear to condition SPE cross-linguistically. These constraints themselves involve a collusion of factors related to domain-general cognitive processes, language-specific reflexes, and other aspects of discourse organization. These constraints include priming, anaphoric distance, discourse organization (prosodic and syntactic linking, inter-clausal configurations, clause type, morphosyntactic status of the antecedent), and

general antecedent accessibility (which implicates all of the above constraints as well as the semantics of discourse referents).

4.6 TAKING STOCK: SUBJECT PRONOUN EXPRESSION/ANAPHORA RESOLUTION IN CABO-VERDEAN CREOLE.

In this chapter, we have seen that CVC has a disjunctive tonic-atonic subject pronoun inventory with SCs, and X2SBJs that involve a lexical DP and a SC. Some researchers have argued for a tripartite classification based on theory-internal claims about the formal properties of strong, weak, and clitic pronouns in Cardinaletti's & Starke's (1994, 1996, 1999) Typology of Structural Deficiency (cf. Baptista 2002; Pratas 2004). While it is not an objective of this volume to comprehensively test the claims implied by the tripartite classification, the semantic properties associated with pronominals and their assumed relationship to antecedents under the Typology of Structural Deficiency will be important.

Since diachronic perspectives on SCs assume that they come about by a process of grammaticalization that acts upon independent pronouns, there has been much dispute in several languages and across theoretical paradigms over the pronominal vs. affixal status of SCs, or the kinds of agreement or anaphoric relationships in which they can engage. A primary goal of this study is to approach this issue using a different perspective: by examining the variable rule system underlying the distribution of anaphoric elements that compete with SCs in the CVC subject domain, namely X2SBJs and ∅ subjects.

The distribution of ∅ subjects and the role of CVC with respect to the NSP have been a matter of dispute (cf. Baptista 1995, 2002; Pratas 2004; Costa & Pratas 2008, 2013). Following Baptista (2002), and based on data in my own corpus, I maintain that argumental anaphoric ∅ subjects are possible in CVC, albeit they occurred at low rates (several examples are exhibited in the following chapter). In BP, inanimate and nonspecific referents have been found to promote ∅ subjects (cf. Kato & Duarte 2003, 2005; Duarte & Soares da Silva 2016). Based on the notion that

pronouns like (∅ morpheme) *pro* and SCs can bear inanimate, indefinite, and nonspecific reference precisely because they are 'referentially deficient' (cf. Cardinaletti & Starke 1994, 1996, 1999), Kato & Duarte (2003, 2005) proposed an "avoid referentially deficient pronoun" constraint to explain patterns of ∅ subjects in BP. Since in CVC, like in other Lusophone-origin vernaculars, the same set of overt pronominals compete with ∅ subjects to resume inanimate, indefinite, and nonspecific antecedents (cf. 87-88, above), it is likely such a constraint is active in CVC too.

The extent to which indefinite and nonspecific lexical DPs can participate in X2SBJs has been taken to be indicative of the degree of grammaticalization associated with SCs cross-dialectally in Gallo-Romance and Northern Italian Dialects (cf. Culbertson 2010). A diachronic increase in use of the topic-comment discourse organization strategy resulting in frequent reliance on left-dislocated lexical DPs in topic-shift constructions has been proposed as a catalyst for the grammaticalization of independent subject pronouns into atonic SCs under 'The NP Detachment Hypothesis' (Givón 1976). Culbertson (2010) considers this to be the process driving increased usage of X2SBJs in colloquial registers and regional varieties of French, and many researchers of BP have made note of the increase in X2SBJs/topic-comment structures diachronically in BP (Pontes 1987; Negrão & Viotti 2000; Modesto 2008; Oliveira da Silva & Álvez Fonseca 2018), something that has been attributed to BP's transformation from a consistent to a partial NSL (Duarte 1993, 1995; Barbosa, Kato, & Duarte 2005). Tavares Silva, Carvalho, & Ziober (2018), observe that indefinite lexical DPs can participate in X2SBJs in BP at low rates. Like Culbertson for Northern Italian Dialects and Gallo-Romance, I took this to be indicative of the grammaticalization of subject pronouns in BP towards SC status; the clitic status of BP subject pronouns has been argued for by Luís & Kaiser (2016:230) independently from the facts about X2SBJs in BP.

Importantly, both French and BP have historically shifted away from consistent NSLs (cf. Kaiser 2009), although the grammaticalization on French SCs is clearly much greater than BP subject pronouns. Since we know that CVC's superstrate language was a consistent NSL, and since CVC, like French and BP historically, experienced an intertwined reorganization of its subject

pronoun inventory and loss of suffixal subject-verb person-number agreement morphology from the superstrate, we might assume that the development of SCs in CVC has its origin those processes of diachronic language change. Alternatively, since there is strong evidence that the CVC 'bare' verb is based on the superstrate infinitive (Quint 2008b), the absence of person-number inflection in CVC would not have come about by erosion, but by a generalization of the lack of number agreement on superstrate nonfinite forms. Whatever the diachronic reason for the lack of person-number inflection in CVC may be, the consequence would have been the initiation of a grammaticalization process on existing pronominals. Wratil's (2011) 'Null Subject Cycle' proposes a recapitulation of the NSP that accounts for the grammaticalization of pronominals, the erosion of bound inflection, the reintroduction and recruitment of new pronominals, and how these interact to modulate the degree to which a language is a partial NSL.

Finally, the distribution of SPE in CVC is likely beholden to the same set of predictive constraints found to condition SPE cross-linguistically in Probabilistic Grammar-oriented studies (Claes 2017). Some of these include: the syntactic role and morphological composition of the antecedent (PRIMING, ANTECEDENT ACCESSIBILITY PATTERN, PERSON/NUMBER, SURFACE MORPHOLOGICAL FORM OF THE ANTECEDENT) and the clausal and coreferential relationship between anaphor and antecedent (ANAPHORIC DISTANCE, ANTECEDENT ACCESSIBILITY PATTERN, CLAUSE TYPE, SWITCH REFERENCE). These constraints, along with those related to issues of SPE from the prior literature discussed throughout this chapter, inform the methodological perspective and experimental design of the present study, which brings me to the topic of the next chapter.

Chapter 5: Methodology

In the previous chapters, we have seen that Subject Pronoun Expression (SPE) in Cabo-Verdean Creole (CVC) has been explored extensively from a Generative Grammar perspective (Baptista 2002; Pratas 2004) and from a historical perspective (Quint 2000a; Lang 2012). The present study is the first to approach the issue by combining a quantitative 'Probabilistic Linguistics' (Claes 2017) methodology with theoretical constructs for Generative Grammar (GG) approaches. The motivation for adopting this approach is to bring new evidence to bear on protracted debates over the morphosyntactic status and clausal position of subject clitics (SCs) and null subject/zero anaphora (Ø) in CVC and cross-linguistically. Previous research on anaphoric referring expressions like SCs and Ø subjects have depended largely on the *a priori* theoretical assumptions of the researcher; meanwhile, classical 'tests' for clitichood (e.g. Kayne 1975; Zwicky & Pullum 1983; *inter alia*) have proven incapable of resolving the persistent question of whether SCs are independent subject pronominals or bound inflectional affixes.

Research programs that explore Ø anaphora have also diverged on theory-internal grounds; for instance, those exploring the typology of Ø subjects under the Null Subject Parameter (NSP) have not always engaged with researchers investigating Ø anaphora in the 'Probabilistic Grammar' (Claes 2017) traditions (e.g. usage-based approaches, functionalism, 'cognitive linguistics', typology, variationist sociolinguistics, etc.). The application of a variationist sociolinguistic methodology to explore issues of SPE in languages like CVC may help to move past the theoretical 'standoffs' over the formal and functional status of these various subject anaphora; for example, instead of relying of 'classic' tests for clitichood, inferential statistical modeling can be used to compare the distribution of SCs against other referring expression that compete with SCs for nominative anaphora resolution.

I begin this chapter by recapitulating the theoretical and methodological perspective guiding this study (see also Section 1.3) in Section 5.1. Next, in Section 5.2, I revisit the main themes of this volume, and in Section 5.3, I enumerate specific hypotheses and predictions related to the broad themes discussed in 5.2. Since the this is the first study to apply a variationist sociolinguistic methodology to the issue of SPE in CVC, I dedicated Section 5.4 to a delimitation of the variable context or envelope of variation. After that, I describe the transcription procedure and methods for coding the independent predictors in Section 5.5; since the transcription procedure directly informs the way in which the predictor variable was coded, I first described the transcription procedure in 5.5.1, and then explain the methods for coding each of the independent variables in 5.5.2. In Section 5.6, I provide a description of the speakers in the sample and the locales for data collection. Finally, In Section 5.7, an overview of the statistical procedures is provided.

5.1 A RECAPITULATION OF THE THEORETICAL PERSPECTIVE IN THIS DISSERTATION.

The methodological and theoretical approach adopted in the present study draw on findings from rationalist as well as empiricist traditions. I adopt the view that the synthesis of these approaches will ultimately lead to the most parsimonious models of Language, not only in terms of descriptive (typological) adequacy, but also explanatory adequacy with respect to the cognitive aspect of the human language capacity, which like so many other human faculties, is likely governed by an innate component that interacts with a statistical learning component, and by language-specified domains that interface with domain-general process, all of which themselves are modulated by social and behavioral phenomena (Harris 2006; Culbertson & Kirby 2016).

This is not wholly novel, I follow Tarallo (2015[1987]:54), Tarallo & Kato (2007[1989]), Martínez Sanz (2011), and Duarte & Soares da Silva (2016:4) in the specific application of this epistemological perspective to study design. Some of the predictor variables in the present study are based on formal theoretical constructs (e.g. anaphor binding and the categories in the Typology

of Structural Deficiency), but also constraints of concern in the 'Probabilistic Grammar' (Claes 2017) literature, such as those related to discourse/pragmatic organization, online processing, and sociocultural indexation. Just as findings in the GG tradition are considered when selecting constraints for testing in the variationist model, the fruits of the quantitative portion of the investigation can also 'feedback' into theoretical paradigms from the GG tradition to help improve formal modeling, particularly with respect issues of language change, the role of domain-general (rather than language-specific) cognitive processes, and other language-external forces that similarly impact the outcomes of linguistic variation. Duarte & Soares da Silva (2016:4) described such an approach:

> […] empiricism and rationalism could be complementary and that to consider the internal factors so as to understand the source of variation could help the analyst to infer more general principles of grammar. [...] the properties usually associated with parameters, as proposed in Chomsky (1981), contribute to formulating hypotheses and specifying the factor groups as a first step to investigate the variation of forms in competition that necessarily precedes any linguistic change; in the same way, the tendencies identified by variationist research help determine, revise, and update the properties related to the parameter issue.

Along these same lines, Martínez Sanz (2011:246), in exploring linguistic restructuring in Dominican Spanish SPE, outlines a similar methodological/theoretical synthesis:

> The fact that [formal] structural properties arise as particularly relevant for disentangling grammatical restructuring in subject expression argues, in my view, in favor of applying a cohesive approach to syntactic variation to arrive at an adequate description of Caribbean subject grammars. The available theoretical investigations can inform quantitative research on the structural variables that might potentially regulate these grammars and that have been overlooked in previous variationist work. In turn, taking a corpus (to which variationist methodology is applied) as a point of departure to draw theoretical conclusions will allow syntactic investigations to build accurate theories on the possibilities of subject expression afforded by Caribbean grammar. […] It is the stand taken in the present study that to accurately define these potential sites of variability is a necessary step to build syntactic theories that can explain the limits afforded by formal grammars for cross-linguistic and cross-dialectal variation.

In addition to demanding that the rationalism-oriented researcher consider matters of variation and probability in language, it also requires that the quantitative-oriented researcher take into account abstract considerations of clause structure, binding relationships, and the role of

underlying hierarchical structure. While I assert that the synthesis of these principles is relevant for the study of any language, and Language generally, it is particularly relevant for examining SPE in CVC, as the brunt of the previous literature on this topic has been developed within a GG framework, or has focused on atheoretical description, documentation, and historical reconstruction.

5.2 A RECAPITULATION OF THE BROAD THEMES IN THIS DISSERTATION.

This subsection revisits the main themes of this volume in anticipation of enumerating specific research questions and hypotheses/predictions in the following subsection. The first recapitulation is stated as a question; the following answer revisits the broad themes identified in the discussion of the previous literature from the prior chapter:

(i) Based on previous research, what are the morphosyntactic, semantic-referential, discursive, and diachronic forces that are relevant for Subject Pronoun Expression in Cabo-Verdean Creole?

We have seen competing classifications of overt subject pronouns in CVC that reflect the theoretical orientations of the researcher. Those working outside the Generative Grammar (GG henceforth) tradition have adopted a bipartite analysis based on a disjunctive tonic-atonic opposition (Veiga 1995, 2002; Quint 2000a,b; Lang 2012). Those working within the GG tradition have adopted the tripartite analysis, thus advancing the claim that CVC lexicalizes all three universal classes of pronominal under the Typology of Structural Deficiency (Cardinaletti & Starke 1994, 1996 1999): strong, weak, and clitic (Baptista 2002; Pratas 2004). The properties associated with each class are argued to be the consequence of "structural deficiency", literal differences in the functional projections underlying each one (Cardinaletti & Starke 1994, 1996, 1999). Resulting from this structural deficiency are a set of morphosyntactic, semantic, and

prosodic properties; the classes form a proper subset with respect to asymmetries in their adherence to formal features: *strong* ≥ *weak* ≥ *clitic*. The semantic properties associated with referential deficiency are of particular importance for the present study; deficient pronominals can bear inanimate, indefinite, and nonspecific reference. Thus, like in other Lusophone-based vernaculars, the same set of overt pronominals are used for referentially deficient and fully referential specified antecedents alike. The relationship between Ø subjects and semantically referentially deficient antecedents is consistent with the claims that *pro* is a deficient pronoun (Cardinaletti & Starke 1994:68,89-91) (see Sections 4.1 and 4.3).

As in other languages that have both atonic SCs and tonic subject pronouns, the status of the former is contested. Under the 'inflectional affix hypothesis', SCs are assumed to be tense heads or bound inflectional affixes, rather than nominative pronominals in a subject position. On the other hand, the 'phonological clitic hypothesis' takes SCs to be genuine nominative pronominal arguments of the verb that cliticize only at the phonological level (cf. Kayne 1975; Brandi & Cordin 1988; DeGraff 1993; Déprez 1994; Baptsita 1997, 2002; Poletto 2000b; Pratas 2005; De Cat 2005; Culbertson 2010; *inter alia*) (see Section 4.2).

Similar debates are reflected in the Probabilistic Grammar (Claes 2017) literature. While some languages are thought to have unambiguous anaphoric person/subject markers (~pronominal phonological clitics) that engage in non-local anaphoric agreement, others are thought to have grammatical person/subject markers (~bound inflection affixes marking subject-verb agreement) that engage in local grammatical agreement. Others still have ambiguous person/subject markers; these can engage in grammatical agreement in a local domain or anaphoric agreement in a non-local domain (cf. Bresnan & Mchombo 1987; Siewierska 2004:121-126; Kari 2017; *inter alia*) (see Section 4.4).

Crucially, a cross-linguistically common grammaticalization cline involves the erosion of independent pronominals into atonic SCs/person markers, then into bound inflectional affixes, which themselves can be eroded into Ø (Givón 1976; Siewierska 2004:22; Wratil 2011) (examples

49 and 50 in 4.4, above; repeated as examples 102 and 103, below, respectively) (see also Hopper & Traugott 1993:7; Campbell & Harris 1995:20).

(102) a.) demonstrative pronoun → stressed independent pronouns
b.) stressed independent pronouns → unstressed clitic pronouns
c.) unstressed clitic pronouns → obligatory verb agreement

(103) weak > clitic > bound > zero

Under the 'NP Detachment Hypothesis' (Givón 1976), one catalyst for this grammaticalization process is increased reliance on topic-shifted structures that involve the left-dislocation of a lexical DP (example 51, Section 4.4, above, repeated below as example 104).

(104) *Amilson, E ta kumi txeu dimás*
A. […] 3.SG.CL TMA eat a lot too much
'A., he eats way too much.

As this topic-shifting strategy becomes more frequent, its discursive/pragmatic value becomes neutralized, and the independent subject pronoun starts to get reanalyzed as dependent SCs, and eventually bound affixes, yielding the "true doubling structure" (cf. Culbertson 2010) (example 52, Section 4.4, above, repeated below as example 105 below).

(105) *Amilson E=ta kumi txeu dimás*
A. 3.SG.CL=TMA eat a lot too much
'A. he-eats way too much.'

Culbertson (2010) has argued that this process may have contributed to the grammaticalization of SCs in Colloquial French. Some evidence for this was gleaned from the prosodic contours of lexical DPs involved in double-subject constructions (X2SBJs henceforth), as well as the token frequency of X2SBJs in corpora. She demonstrated that lexical DPs in X2SBJs show less prosodic signs of left-dislocation, and overall, X2SBJs are becoming more frequent; both pieces of evidence suggest that Colloquial French SCs are becoming more affix-like. When French X2SBJs are situated typologically among X2SBJs in Northern Italian Dialects and other Gallo-Romance vernaculars, the semantic properties of the lexical DPs involved in the X2SBJ constructions reveal differing degrees of grammaticalization on SCs cross-dialectally (Table 23 in Section 4.2, above). This typology of X2SBJs reveals that Colloquial French SCs are at middle stage in the grammaticalization cline when compared to the more grammaticalized SCs observable in other Gallo-Romance and Northern Italian Dialects. Since in CVC tonic subject pronouns frequently double with SCs, this might be taken as evidence that SCs have lost at least some of their pronominal status.

We also saw evidence that, diachronically, Brazilian Portuguese has developed topic-comment discursive strategies similar to those found in discourse-oriented languages (Pontes 1987; Negrão & Viotti 2000; Modesto 2008 Oliveira da Silva & Álvez Fonseca 2018) and that X2SBJs have become more productive and can even include indefinite lexical DPs (Tavares Silva, Carvalho, & Ziober 2018). Luís & Kaiser (2016) have argued that Brazilian Portuguese subject pronouns are becoming more SC-like. Although CVC SCs are doubtless more grammaticalized than BP subject pronouns, and therefore more 'clitic-like' (or more 'affix-like'), the diachronic changes in BP may lend clues as to the positioning of each language's subject domains by the degree of grammaticalization on subject pronominals (regardless of the diachronic process responsible for erosion or loss of person-number inflection).

This is because diachronic changes in both French and Brazilian Portuguese are related to historical restructuring of the subject and verbal domains. The grammaticalization of Colloquial French SCs is part of much longer sequence of diachronic changes in the history of Gallo-Romance.

Old French was a consistent Null Subject Language with rich inflectional morphology. During the transition to Middle French much of this inflectional morphology was eroded, Null Subject Language (NSL) properties were lost, and SCs began to develop from the grammaticalization of independent tonic pronominals (Kaiser 2009). Though less advanced in this process than French, similar historical reorganizations occurred in the Brazilian Portuguese subject and verbal domains (see Figure 6, Tables 25 and 26, Section 4.5.1), and those reorganizations triggered changes in BP's Null Subject Language status (cf. Duarte 1993, 1995; Barbosa, Kato, & Duarte 2005; *inter alia*) (see Sections 4.5.1 and 4.5.2). In turn, this might be the reason for the increasing SC-like status of Brazilian Portuguese subject pronouns. Wratil's (2011) 'Null Subject Cycle' (see Figure 3, Section 4.2) proposes a recapitulation of the Null Subject Parameter that accounts for the grammaticalization of pronominals, the erosion of bound inflection, the reintroduction and recruitment of new pronominals, and how these interact to modulate the degree to which a language is a partial NSL.

Since CVC 'bare' verbs appear to be based on the superstrate's infinitive (Quint 2008b), person-number inflection would have been lost by generalization of nonfinite forms rather than erosion or restructuring of verbal morphology. As we saw from Lang's (2012) reconstruction of the subject pronoun system of proto-CVC, new SCs came about by cross-linguistic convergence between existing atonic person markers in the substrate languages and reduced forms resulting from a series of changes on superstrate tonic pronouns (likely related to grammaticalization), and by processes of analogical change. We have noted that, when used as nominative arguments, CVC tonic pronouns are almost always deployed in X2SBJ constructions (excluding with copular *e* 'to be'); this is likely indicative of the ongoing grammaticalization of SCs in CVC.

In Probabilistic Grammar (Claes 2017) approaches, SCs and Ø subjects are understood to be "high referentially continuity devices" that are associated with highly activated, salient, or accessible antecedents (Givón 1976, 1983b, 2001[1984], 2017; Ariel 1990). Highly accessible antecedents are associated with a number of properties that have been categorized in accessibility hierarchies like (example 101, Section 4.5.3, above; repeated as example 107, below). SCs compete

with another "high referential continuity device", anaphoric Ø, to resume antecedent with the properties to the left of > in the accessibility hierarchies Siewierska (2004:149).

(106) a.) the person hierarchy

1ST > 2ND > 3RD

b.) the nominal hierarchy

pronoun > noun

c.) the animacy hierarchy

human > animate > inanimate > abstract

d.) the referential hierarchy

definite > indefinite specific > nonspecific

e.) the focus hierarchy

not in focus > in focus

Additionally, SCs will also complete with Ø anaphora at other sites of maximal referential continuity like in clause chain internal positions, though SCs should be favored over Ø anaphora at clause chain boundaries and in switch-reference contexts (cf. Givón 2017:6-7) (see Section 4.5.3).

With respect to the broad-theme research question in (i), the distribution of CVC SCs is likely to reflect the degree to which they have advanced along the pronoun-to-affix grammaticalization cline (102)-(103). This means that SCs might be expected to demonstrate some of the properties of pronouns, but other properties associated with affixes. We have already seen that this is true to a certain extent, since clitichood tests support either a pronominal or affixal analysis. Doubtless, SCs will be the most frequent subject type; they are the primary means for overtly marking the person-number indexation of discourse referents. The discursive, semantic, and morphosyntactic properties of SCs (and their antecedents) will be inferred from the variable rule system that conditions the selection of anaphora that compete against SCs for the subject 'space'. SCs are predicted to be the 'default' or 'discursively neutral' subject type. If the Typology of Structural Deficiency is on the right track, we can also expect to see SCs serving as anaphora for

a range of antecedent types, including those with referentially deficient semantic properties like inanimacy, indefiniteness, and nonspecificity. Since X2SBJs are the primary way in which tonic pronouns are deployed in CVC (except with copular *e*), we might expect X2SBJs to exhibit some of the properties usually attributed to singleton tonic pronouns, like serving as switch reference devices and reintroducing distant discourse referents.

The status of Ø subjects in CVC is also contested. In one sense, this is directly related to the analysis attributed to SCs. Under GG approaches, if the phonological clitic hypothesis is adopted and SCs are analyzed as genuine subject pronouns, they must occupy the canonical subject position (SPEC,IP). Any tonic pronouns participating in a X2SBJ constructions with SCs are thus analyzed as occupying a left-dislocated position, leaving no need to propose an empty category (such as *pro*) in the subject position, since that position is already filled with overt material (see example 34a, Section 4.2; and examples 76-77, Section 4.5.2, above). On the other hand, if the inflectional affix hypothesis is adopted, then SCs are analyzed as bound inflectional affixes; this view suggests that tonic pronouns participating in X2SBJs with SCs are in the canonical subject position. In the absence of a tonic nominative element, then, the subject position must be assumed to filled with some empty category (like *pro*) (see example 34b, Section 4.2; and examples 78-79, Section 4.5.2, above).

Pratas (2004) and Costa & Pratas (2008) tacitly assume the phonological clitic hypothesis and reject Baptista's (2002) examples of Ø subjects (with no overt nominal elements whatsoever). Thus, for them, CVC must be a partial Null Subject Language; root Ø subjects are disallowed, impersonal and generic subject can be expressed as Ø, expletive and other non-argumental subjects are categorically Ø, and anaphoric Ø is restricted to highly specific embedded contexts. Baptista (1995, 2002), on the other hand, understands CVC to be a consistent NSL; this is because she adopts the inflectional affix hypothesis, but also because she found numerous examples of anaphoric Ø subjects in root contexts (where no other overt subject elements are present whatsoever; cf. examples 80-88, Section 4.5.2). I confirm in my corpus the use of such anaphoric Ø subjects.

'Probabilistic Grammar' (Claes 2017) approaches have explored the use of Ø anaphora widely, even in languages that have traditionally been considered non-NSLs under the NSP. These approaches have provided two valuable contributions to the study of Ø subjects that are important for CVC: (i) the same set of predictive constraints appear to condition the realization of Ø subjects

cross-linguistically (no matter the classification of the language under the Null Subject Parameter, nor overall rates of Ø); (ii) Ø subjects are most likely to arise when a combination of predictive constraints come together to maximize antecedent accessibility. However, there may be a property of Ø subjects that is specific to Lusophone-based vernaculars in which semantically referentially deficient antecedents promote the realization of a Ø subject[58]. This is consistent with the claims of the Typology of Structural Deficiency in which *pro* is considered a deficient pronoun and is thus associated with the semantics of referential deficiency like inanimacy, nonspecificity, and indefiniteness. If Ø subjects in CVC can be found to tend towards these semantic properties, an extension of Kato's & Duarte's (2003, 2005) "avoid referentially deficient pronouns" constraint might be applied to CVC, given that this constraint is taken to be probabilistic in nature.

(2) How can we situate CVC with respect to theoretical models that attempt to delineate languages by the organizational properties of the subject domain (e.g. the Null Subject Parameter, the Typology of Structural Deficiency, the Null Subject Cycle)?

Given that this dissertation adopts a quantitative analysis (with certain predictive constraints based on theoretical constructs in formal approaches) to examine SPE in CVC, any of the properties attributed to subject anaphora must necessarily be understood to be probabilistic in nature. Since the realization of Ø subjects in inflection-'poor' languages like CVC usually depends on a confluence of several constraints, rather than some constant formal criterion, when and where Ø subjects are used will necessarily be variable and contingent upon the degree to which the relevant constraints coincide and are instantiated in a given context. Further, since elements like SCS are by their nature morphemes at the middle stage of a grammaticalization cline, their associated properties are likely to evade simplistic categorical classification.

These facts may have posed a problem for Principle & Parameters approaches that were too rigid in their interpretations of cross-linguistic variation. However, microparametric

58 Alternatively, this may simply be a property of languages where the same set of overt pronouns are used to resume referentially deficient and referentially fully specified antecedents alike; the probability of selecting a Ø subject in these languages may become higher with semantically referentially deficient antecedents.

approaches that view the Null Subject Parameter as applying in a gradient rather than categorical fashion (e.g. Duarte & Soares da Silva 2016) may be capable of modeling the probabilistic nature of Ø subject use, the behavior of 'in-between' morphemes like SCs, and the uncertain classification of CVC among Null Subject Languages (NSLs). The Minimal Morphological Threshold (Cole 2009; Camacho 2011, 2013) explains how Ø subjects are identified by antecedents even in languages where inflectional morphology is usually responsible for identification, and that different languages set different criteria for how identification of Ø subjects is carried out. If Baptista's (1995, 2002) analysis of CVC as a Null Subject Language is assumed, then SCs are person-number inflection and serve to identify Ø in the subject position, while true anaphoric Ø is identified by a discourse antecedent when SCs are dropped. Similarly, Wratil's (2011) Null Subject Cycle captures diachronic interactions between the nominal and verbal domains in which the erosion of bound inflection can drive the grammaticalization of tonic pronouns into SCs. If instead person-number inflection in CVC was lost because the CVC verb is based on the superstrate infinitive (Quint 2008b), the grammaticalization cline responsible for turning independent pronouns into SCs would nonetheless have been initiated during the early restructuring phase of this language.

New improvements to the Minimalist Program have also found ways to account for cross-linguistic variation via the valuation of formal features stored in the lexicon (cf. Wilson & Henry 1998; Adger & Smith 2003, 2005; Cornips & Corrigan 2005a,b; Sessarego 2011, 2019; *inter alia*). The Typology of Structural Deficiency already allows for some flexibility in its proposed categories, for instance, with respect to asymmetries in their various properties:

> […] while these asymmetries seem to be universal, none of the interpretive asymmetries is systematic: It is not the case that there is a strict covariation between being of one class, and having one type of semantic/phonetic interpretation. The interpretational characteristics are asymmetric but overlapping: the three classes are purely abstract (for example, both deficient and strong elements can refer to human entities and to prominent discourse referents, although an asymmetry holds with respect to non-human entities and non-prominent referents. (Cardinaletti & Starke 1999:213).

Finally, 'Probabilistic Grammar' (Claes 2017) approaches that rely on the concept of antecedent accessibility may need some way to explain the relationship between Ø subjects and referentially deficient antecedents in languages like BP (and possibly CVC), since semantic properties like inanimacy, indefiniteness, and nonspecificity are assumed to make discourse referents less salient, and therefore less accessible for resumption by Ø anaphora (cf. Ariel 1990).

5.3 HYPOTHESES AND PREDICTIONS.

In this section, I will describe the hypotheses and predictions associated with the broad-theme questions from Section 5.2. These include linguistic-structural and language-internal factors 5.3.1., among them, the overall rates of the anaphoric elements involved in SPE in CVC 5.3.1.1., constraints related to discourse organization, the relationships between clauses, and issues of coreference/binding/anaphora resolution 5.3.1.2, variables concerning the antecedent and its semantic-referential value 5.3.1.3., and predictors involving the verb phrase 5.3.1.4.

In 5.3.2., I consider hypotheses and predictions related to language-external and sociological factors, many of which are treated as exploratory and pending empirical confirmation given the lack of prior examination of the correlations between sociocultural categories and SPE in CVC. In 5.3.2.1., factors related to the data collection materials and the identity of the interviewer(s) are discussed. Possible individual-specific sociocultural constraints are addressed in 5.3.2.2. Before proceeding to discussion of the each of the research questions to be posed over the course of the next section, I list each of them for referential purposes and clarity:

RQ1: At what rates do the three nominative anaphoric elements in CVC (SCs, Ø, X2SBJs) occur?

RQ2: What are the constraints related to discourse organization, coreference, and referent tracking, that influence SPE in CVC? What is the nature of the effect?

RQ3: What are the constraints related to the person-number, syntactic role, and semantic-referential properties of the antecedent, that influence SPE in CVC? What is the nature of the effect?

RQ4: What are the constraints related to the verb phrase that influence SPE in CVC? What is the nature of the effect?

RQ5: What are the language-external non-individual-specific constraints that influence SPE in CVC? What is the nature of the effect?

RQ6: What are the language-external individual-specific factors that influence SPE in CVC? What is the nature of the effect?

5.3.1 Language-internal/Linguistic-structural hypotheses.

This section describes the hypotheses and predictions associated with language-internal constraints and their effect on variable SPE in CVC. It was established that these are informed by variationist sociolinguistic studies of SPE in Portuguese, Spanish, and English, by theories on Ø subjects, clitics, and the typology of subject domains developed within the GG tradition, and by 'Probabilistic Grammar' (Claes 2017) approaches to atonic pronominals and anaphoric Ø. Before any consideration of these constraints, however, the relevance, if any, of rates of SPE need to be considered.

5.3.1.1 Rates of overt and covert subject expression.

To a certain extent, rates of SPE have been taken to be indicative of partial NSL status (cf. Duarte 1993, 1995; Barbosa, Kato & Duarte 2005; Duarte & Soares da Silva 2016). However, rates alone are not enough; we have seen that the NSP is associated with a cluster of properties, and partial NSLs have also been found to be losing some of the properties in the NSP cluster (cf. Duarte 1993, 1995; Toribio 1993b, 2000, Kato & Duarte 2003, 2005; Barbosa, Kato & Duarte

2005; Holmberg *et al.* 2009; Camacho 2012b, 2016; *inter alia*). Further, rates of (non-)expression may differ from language to language depending on person-number configuration, clausal status, and other considerations of discourse (cf. Vainikka & Levy 1999; Wratil 2009, 2011; Camacho 2013).

Variationist studies have also observed that overall rates of SPE can differ greatly across languages (Posio 2012, Torres Cacoullos & Travis 2019), and across varieties of the same language (c.f. Carvalho, Orozco, & Lapidus Shin eds. 2015). Despite variation in overall rates, the variable rule systems underlying SPE are often sensitive to a similar set of probabilistic constraints across languages, although the degree of the effects and the ranking of levels within the constraints will often differ cross-linguistically and -dialectally (cf. Otheguy, Zentella, & Livert 2007, and references therein; Martínez Sanz 2011, and references therein; Otheguy & Zentella 2012, and references therein; Carvalho, Orozco, & Lapidus Shin eds. 2015, and references therein). It is for this reason that Torres-Cacoullos & Travis (2019:2-3) assert that: "Rates of use are thus equivocal, being susceptible to the preponderance or dearth in a data set of some propitious context, which may be fortuitous or due to extra-grammatical, situational considerations, such as register or topic." At the same time, Duarte & Soares da Silva (2016) have shown that gradient proportions of covert/overt SPE across Romance varieties correlate with differential effects for independent predictors such as ANTECEDENT ACCESSIBILITY PATTERN, PERSON/NUMBER, and the referential semantics of subject pronominals. Camacho (2016) also ties changes in overall rates of SPE to broader modifications in the identification mechanisms underlying diachronic changes in the NSP status of partial NSLs like Brazilian Portuguese and Dominican Spanish.

Thus, there may be at least some connection between overall rates of covert/overt SPE and the NSP properties demonstrated by a language. Fluctuations in rates of SPE in apparent time may also be indicative of a change in progress in the status of a language under the NSP (cf. Figure 7, from Duarte & Soares da Silva 2016; Section 4.5.3.). However, rates need to be considered in the context of the broader variable system and the predictive constraints that comprise it.

RQ1**:** At what rates do the three nominative anaphoric elements in CVC occur (SCs, Ø, X2SBJs) occur?

H1: In CVC's three-way system of SPE, the outcomes are a double-subject construction (X2SBJ), a subject clitic (SC), and a null subject/anaphoric zero (Ø). SCs are expected to be by far the most frequent subject form. If SCs are phonological clitics, the reason for their productivity might be attributed to a lack of person-number morphology on the verb. If SCs are inflectional affixes, it is due to their role in marking person-number agreement. Under the view that SCs are ambiguous person agreement markers, their high frequency would be explained by their ability to carry out either local grammatical agreement or non-local anaphoric agreement (cf. Siewierska 2004). It is predicted that, due to their grammaticalized nature SCs will exhibit some the features associated with pronominals, and other features associated with affixes, most likely aligning with an analysis in which SCs are akin to ambiguous person markers.

Rates of X2SBJs are expected to be low, similar to the rates of overt tonic subject pronouns in consistent NSLs, as is contingent on their person-number configuration. Similarly, X2SBJs are expected in the same discursive/pragmatic contexts where overt subject pronouns have generally been thought to occur in consistent NSLs: to mark switch-reference and reintroduce distant discourse referents (Givón 2017), contrastively (Mayol 2010), or emphatically (Lezama & Almor 2011).

Anaphoric Ø is predicted to be the least frequent CVC subject type. Ø is most likely to be realized when a confluence of predictors conspire to maximize antecedent accessibility. The predictors will be explored in more detail in the following subsections, but include contexts when the antecedent: is the subject of the immediately prior clause (accessibility), was also Ø (priming), and is proximate to its coreferential anaphor (anaphoric distance). Antecedent accessibility is usually

thought to be maximal with highly salient antecedents, for example, those with semantic-referential properties like definiteness, animacy, and specificity (cf. Ariel 1990), and with 1st and 2nd person referents (Speech Act Participant pronouns) (Wratil 2011).

However, in Lusophone vernaculars, where overt pronouns compete with Ø to resume referentially fully specified and referentially deficient antecedent alike, Ø may be favored with inanimate, indefinite, and nonspecific antecedents (cf. Kato & Duarte 2003, 2005; Duarte & Soares da Silva 2016). Furthermore, Ø subjects in Brazilian Portuguese are most frequent in the 3rd person (Duarte & Soares da Silva 2016). Whether these properties can be associated with Ø anaphora in CVC as well is a matter pending empirical confirmation, but since CVC already employs Ø frequently with generic and impersonal constructions, it would not be surprising to find that Ø subjects are realized when there is an indefinite or nonspecific lexical DP antecedent. Baptista (2002:259) provided one example of anaphoric Ø with an inanimate antecedent (example 87, Section 4.5.2), and she also argued for an 'abstract 3rd person features' identification mechanism for Ø in CVC (Section 4.5.2).

5.3.1.2 Constraints related to discourse organization, clausal linking, coreference, and referent tracking.

This subsection brings together considerations of coreference and pronominal binding, with issues of discourse/pragmatic organization, clausal coherence, and domain-general cognitive processes that intervene in nominative anaphoric relationships.

RQ2: What are the constraints related to discourse organization, coreference, and referent tracking, that influence SPE in CVC? What is the nature of the effect?

H2: Under a variety of theoretical and methodological perspectives, the issue of antecedent

accessibility has been found to be among the most important factors conditioning SPE cross-linguistically. Antecedent accessibility accounts for coreference/anaphora resolution/binding relationships, the syntactic role and morphological composition of antecedents in their containing clauses, the clausal configuration between anaphor and antecedent, and the influence of domain-general cognitive constraints such as the anaphoric distance between a target and its antecedent and linguistic structural priming.

Since SCs are assumed to be the most frequent of the three nominative anaphora, they might be considered a sort of default subject type, pragmatically neutral, and likely to appear across a variety of clausal configurations, anaphoric distances, and other contexts. Therefore, the effects for antecedent accessibility will likely be manifested most obviously in the realization of Ø and X2SBJs, when these are compared to SCs. That said, SCs are most likely to be realized over Ø subjects at clause chain boundaries and in switch-reference contexts.

Duarte & Soares da Silva (2016), in their cross-Romance comparison of variable SPE, found that in the most consistent NSL of the group, Italian, anaphoric distance was the most significant predictor of Ø: Italian Ø subjects were favored across a range of anaphoric distances and in various clausal configurations, including those in which the antecedent was in a non-adjacent clause relative to that of the anaphoric subject. Given high overall rates of Ø in Italian across discourse contexts (i.e. no pragmatically unmotivated overt pronouns), they observe that "even in unfavorable structural conditions the 'avoid pronoun principle' (Chomsky 1981) is active in Italian." (2016:9).

In Romance varieties like European Portuguese and Argentine Spanish, which Duarte & Soares da Silva take to be less consistent in their adherence to the properties of the NSP, short anaphoric distances remain an important predictor, but

the effects of a c-commanding antecedent become highly predictive of Ø (2016:10-17).

In varieties like Brazilian Portuguese and Puerto Rican Spanish that have shifted away from NSL status and are best classified as partial NSLs, Duarte & Soares da Silva take the context of a c-commanding antecedent to be a "resistance context", the one antecedent accessibility configuration that still favors Ø (2016:17-19). However, underpinning Duarte's & Soares da Silva's analysis is the notion that a loss in the number of oppositions in the verbal inflection paradigm of BP and PRS is what has driven the erosion of their status under the NSP.

As long as SCs are not taken to be inflectional affixes, it is clear that any loss of person-number agreement inflection in the history of CVC would have been much more extensive than in contemporary varieties of BP or PRS, implicating an entire inflection paradigm, rather than the loss or merger of just one or two categories within the paradigm. Further, the absence of person-number inflection is likely due to the CVC bare verb having been based on the superstrate infinitive (Quint 2008b). It is unclear, then, whether a c-commanding antecedent will be a relevant predictor (perhaps a "resistance context"?) for the occurrence of Ø in CVC.

Indeed, Costa & Pratas (2013) concluded that a c-commanding indefinite or WH-operator is the only context capable of licensing embedded Ø in CVC. However, the many examples of Ø in Baptista (2002) are in simple main clauses. The same is true for the examples of Ø presented in Lipski (1999) for several Iberian-lexifier creoles; although, just as Kouwenberg (1990) argued for Papiamentu, Lipski concluded that these instances of Ø are not consistent with the properties of *pro* due to their non-referential/argumental, or at least ambiguously referential status (i.e. they might be big PRO or pro_{ARB}, hence the possibility of generic or impersonal readings) . Kouwenberg, however, argues for the possibility of anaphoric Ø in embedded *pa*-clauses (example 90, Section 4.5.2, above).

In English, there is evidence for an Intonational Unit-initial constraint on Ø (Torres Cacoullos & Travis 2019) or a preference for Ø in turn-initial first-subject contexts (Wagner 2016). In languages in which Ø competes with tonic and atonic pronouns in SPE, discourse organizing strategies such as clausal linking and chaining tend to favor the realization of Ø and SCs in simple main clauses and in clause chain medial position (Givón 2017), particularly under an analysis where Ø is taken to be a "high referential continuity device" (Givón 2017:284-286). Generally, from a typological perspective, languages differ greatly in their treatment of Ø in embedded contexts (Givón 2017:240-245) (see Section 4.5.3).

Contradictory observations from past studies of CVC leave the relevance of a c-commanding antecedent for SPE a question awaiting empirical confirmation, but clues from other Romance-lexifier creoles (Kouwenberg 1990; Lipski 1999) and other languages with Ø, atonic, and tonic pronouns, indicate that Ø will be more likely in simple main clauses in the medial position of a clause chain.

I hypothesize that - given the degree to which CVC differs typologically in its inflection system when compared to partial NSL varieties of Portuguese and Spanish - the selection of Ø in CVC will tend to occur in simple main clauses in chain medial position, or across prosodically linked Intonational Units with coreferential antecedent-anaphor, as in other languages where SCs and tonic pronouns compete with Ø (Givón 1983b; 2017).

This is because, in the absence of a system where robust person-number inflectional morphology encoding person-number agreement plays a role in the identification of Ø subjects, the selection of Ø might be more sensitive to the effects of antecedent accessibility, such as those related to discourse organization (clause chaining and linking, subject reference continuity, other aspects of antecedent accessibility) and domain-general cognitive processes (priming → implicit learning, working memory; anaphoric distance → working memory, attention), rather than formal

devices representing clausal hierarchies and structural embedding (see Section 4.5.3).

Probabilistic Grammar (Claes 2017) accounts of SPE/anaphora resolution have shown that switch-reference contexts favor overt tonic pronouns over Ø subjects, and that tonic pronominals compete with SCs in switch-reference contexts, but that tonic pronouns should be favored over SCs when referents are less accessible or more anaphorically distant in the discourse. Since X2SBJs are the primary way of deploying a tonic pronoun in CVC (except for with copular *e*), switch-reference contexts are predicted to favor X2SBJs, just like with overt tonic pronouns in languages lacking atonic subject pronominals.

Linguistic structural priming, or persistence, has also been shown to be determinative in SPE/anaphora resolution cross-linguistically (cf. Carvalho, Orozco, Lapidus Shin eds. 2015; Wagner 2016; Bouchard 2018; *inter alia*). Priming in SPE entails that the morphological form of an antecedent will trigger repeated uses of those same forms on subsequent anaphora. Therefore, X2SBJs are predicted to prime subsequent X2SBJs, SCs are predicted to prime subsequent SCs, and Ø subjects are predicted to prime subsequent Ø subjects.

While the effects of anaphoric distance are already partially accounted for by antecedent accessibility pattern, when taken as an independent predictor, anaphoric distance has been shown to be highly relevant in determining covert/overt variation for a number linguistic categories (cf. Szmrecsanyi 2006), particularly when measured in number of words (Hinrichs, Szmrecsanyi, Bohmann 2015), and is cross-linguistically determinative for SPE, with average anaphoric distances per subject type falling within a similar range in languages in which Ø competes with tonic and atonic subject pronouns (Givón ed. 1983; Givón ed. 1997; Givón 2017:288-290) (see Section 4.5.3).

Short anaphoric distances are predicted to exert a favoring effect on the realization of Ø, while longer ones will favor the realization of X2SBJs, and SCs are predicted to be distributed across anaphora that lie at a range of distances from their antecedents.

Prosodic and syntactic linking have been shown to be predictive in the selection anaphoric Ø. Torres Cacoullos & Travis (2015, 2016, 2019) have shown that prosodic and syntactic linking across adjacent Intonational Units with coreferential antecedent-anaphor promotes Ø subjects over overt subject pronouns in English and Spanish (see Section 4.5.3).

Linking, which enhances antecedent accessibility, is a good candidate for a constraint that is likely to be relevant for SPE on a general cross-linguistic basis, along with constraints like priming and anaphoric distance. This is likely because linking makes simpler the processing task of tracking referents across a discourse in speech (and thus factors into one's inherent knowledge of SPE). It is possible that these sorts of constraints (which are rooted in domain-general cognitive processes), though active in all languages, may exert an even greater probabilistic effect on the realization of anaphoric Ø in languages that lack inflectional morphology encoding person-number agreement as means of identifying Ø.

I posit that both prosodic and syntactic linking will be predictive of Ø and SCs over X2SBJs, and that when linking interacts with other constraints such as clause chaining, priming, and anaphoric distance, this will exert a favoring effect on the occurrence of Ø over SCs.

The organization of successive clauses into discourse chunks, in which the clauses are unified by thematic, referential, and prosodic continuity, has also been found to effect SPE/anaphora resolution. This is based on the position in which the anaphoric subject occurs in a clause chain (Givón 1983b; 2001[1984]; 2017) (see Section 4.5.3).

SCs and Ø are favored chain-medially, while tonic forms may be favored chain-initially or chain-finally, though these correlations can vary cross-linguistically based on the available inventory of pronominals, head directionality, and other typological properties. Since CVC is almost completely an SVO language, and since X2SBJs are predicted to pattern like overt tonic pronominals, I hypothesize that X2SBJs occur more often chain-initially.

Since Ø and SCs are both expected to abound chain-medially, it is predicted that Ø will be selected over SCs in chain medial position when multiple constraints interact to promote Ø, for instance, if clause-chaining combines with linking and short anaphoric distances.

Related to all these constraints is the notion that high referential continuity promotes anaphoric Ø, followed in probability of realization by SCs. Thus, the most accessible possible condition for an anaphoric subject is: (i) when the antecedent is in an immediately prior clause which is prosodically and syntactically linked to the Intonational Unit containing the anaphor, (ii) when the anaphor is in a clause chain-medial position, and (iii) when a short anaphoric distance lies between the anaphor and its antecedent.

The absence of any of these conditions will be probabilistically less favorable to Ø, and SCs will be more likely to occur. An absence of most or all these conditions will promote X2SBJs.

5.3.1.3 Constraints related to the morphological composition and semantic referential value of the antecedent and anaphor.

This subsection address hypotheses and predictions related to the morphosyntactic composition of the antecedent (person-number, syntactic role in the containing clause) and the antecedent's referential semantic properties (animacy, specificity, definiteness).

RQ3: What are the constraints related to the person-number, syntactic role, and semantic-referential properties of the antecedent, that influence SPE in CVC? What is the nature of the effect?

H3: Person-number has been important in identifying microparametric variation in adherence to the NSP across related languages (Romance) and across dialects of a language (varieties of Portuguese) (Duarte & Soares da Silva 2016). It has also been observed that partial NSLs differ with respect to the person-number configuration in which Ø is deployed (cf. Wratil 2011; Camacho 2013:31-38), and that diachronic changes in the degradation of tonic into atonic pronouns, and then into Ø, can also obtain in different person-number configurations at different times in the history of a language (Poletto 2000; Siewierska 2004; Wratil 2011; Poletto & Tortora 2016). The high topicality of first- and second-person referents (SAP pronouns) are thought to promote Ø due to their 'topic worthiness' and referential salience (Wratil 2011). On the other hand, Baptista (2002) argues for CVC specifically that "default third person features in AGR" allow for the licensing/recovery of Ø. Duarte & Soares da Silva (2016:16) find the highest rates of Ø in BP with 3rd person referents (see also Modesto 2000, 2008).

Thus, past evidence from CVC, and evidence from other related languages, appear to conflict when attempting to predict in what person-number configuration Ø should be most likely in CVC.

Recall that in CVC the individual-level copula *e* 'to be' only admits overt tonic pronouns and Ø (but not SCs) and Baptista (2002:232) suggested that 3rd person subjects of *e* might be a productive context for Ø in CVC. However, nominative arguments of *e* are excluded from the current study since only tonic pronouns and Ø (but not SCs) can serve as nominative pronominal arguments of *e* (see 5.4 ahead on the envelope of variation for SPE in CVC) (see Section 4.5.2).

I posit that, like Brazilian Portuguese, and based on the examples of CVC Ø subjects in Baptista (2002), anaphoric Ø will be most probable with 3rd person antecedents. Once the role of the semantics of 3rd person reference and its influence on SPE is considered later in this subsection, it will become clear why the 3rd person will remain the most productive context for Ø, even when instances of Ø with *e* are excluded.

The syntactic role of the antecedent (e.g. subject, object, oblique) and the antecedent's morphological composition (e.g. pronoun, Ø, DP, demonstrative/deictic, etc.) is also known to affect antecedent accessibility and SPE/anaphora resolution.

Accessibility Hierarchies like those of Givón (1976; 2017:5-7) and Ariel (1990:69-79) identify Ø subjects and atonic pronouns as most accessible or most referentially continuous, followed by tonic pronouns, various types of demonstratives and other lexical DPs. Further, subjects are more accessible to anaphora than objects, and objects are more accessible than obliques (Givón 2017:6). Carminati (2002) has shown for Italian, and Alonso-Ovalle *et al.* (2002) for Spanish, that an anaphoric subject is more likely to resolve as Ø when its antecedent was the subject of its containing clause; an anaphor with an oblique antecedent is more likely to be realized as overt. (see Sections 4.4 and 4.5.3). Therefore, I predict that Ø subjects will be promoted with subject antecedents rather than when an antecedent in some other syntactic role.

The semantic properties of strong, weak, and clitic forms is one of the components of the Typology of Structural Deficiency (Cardinaletti & Starke 1999). Strong pronominals are associated with animate, specific, and definite referents, while weak pronominals and clitics receive their referential properties from their antecedents, and are capable of bearing inanimate, nonspecific, and indefinite reference. The formal category underlying Ø, *pro*, is understood to be a deficient

pronominal. Kato & Duarte (2003, 2005) argue for BP that the "avoid referentially deficient pronouns" constraint is active, since Ø in BP is also promoted by antecedents that bear inanimate and nonspecific reference (see Duarte & Soares da Silva 2016; see also Section 4.3).

I predict that, probabilistically, anaphoric Ø will be favored when its antecedent bears inanimate, nonspecific, or indefinite reference, or any combination of these. Since X2SBJs are the primary way of employing a nominative tonic pronoun in CVC, I assume X2SBJs will exhibit properties similar to those of singleton tonic pronouns. When compared with SCs, X2SBJs should be promoted when their antecedents bear animate, specific, and definite reference. SCs are predicted to be the 'default' subject type and to occur productively with a range of antecedents.

The list for the conditions under which one encounters the 'most accessible possible antecedent' (in 5.3.1.2) can thus be expanded; the most accessible possible condition for an anaphoric subject is: (i) when the antecedent is 3rd person, (ii) when the antecedent is the subject in its clause, (iii) when the antecedent is in an immediately prior clause which is prosodically and syntactically linked to the Intonational Unit containing the anaphor, (iv) when the anaphor is in a clause chain-medial position, and (v) when a short anaphoric distance lies between the anaphor and its antecedent.

The absence of any of these conditions will be probabilistically less favorable to Ø, and SCs will be more likely to occur. An absence of most or all these conditions will promote X2SBJs. A condition that is NOT associated with high antecedent accessibility, but that is nonetheless predicted to favor a Ø subjects in CVC, is when the antecedent is semantically referentially deficient (inanimate, indefinite, nonspecific).

5.3.1.4 Constraints related to the verb phrase.

This subsection addresses hypotheses and predictions related to the verb phrase. The constraints involve overt TMA marking, the overall TMA 'frame' associated with a given proposition, and the lexical class of the verb.

RQ3: What are the constraints related to the verb phrase that influence SPE in CVC? What is the nature of the effect?

H3: There has been little prior research on the relationship between the verb phrase and SPE in CVC. Thus, the hypotheses advanced here are mostly speculative in nature. In English, VP complexity is associated with Ø subjects; the more 'material' in the verb phrase (e.g. modal/auxiliaries, negation, multiple nonfinite verbs, etc.), the more likely one is to realize a Ø subject (Wagner 2016) (see Section 4.5.3). If a similar effect holds for CVC, bare verbs in finite contexts, or the presence of just one TMA particle, should disfavor Ø relative to SCs and X2SBJs. On the other hand, combinations of several TMA particles and/or modals/auxiliaries should favor the realization of Ø over SCs and X2SBJs.

Switches in TMA frame from an antecedent's clause to the anaphor's clause is associated with reduced referential continuity and therefore should disfavor Ø and promote X2SBJs and SCs (for Spanish see Bayley & Pease-Álvarez, 1997; Geeslin & Gudmestad, 2011; see also Section 4.5.3).

The lexical semantics of the main verb may also be deterministic for SPE. In the verb system of languages traditionally classified as creoles, including CVC, the stativity of the verb is relevant for the temporal interpretation of bear verbs (cf. Bickerton 1974:128; Silva 1990). Bare stative verbs usually receive a default nonpast reading while bare non-statives (particularly dynamic or action verbs) usually get a default past reading (see Section 5.5.2.3., below). It is not clear, however, how these lexical verb distinctions would interact with SPE in CVC.

Baptista (2002:255) makes note of Ø with (stative) copular verbs in CVC, but she also provides numerous examples with other verb types.

We can once again append the list of conditions for 'most accessible possible antecedent' (in 5.3.1.3); the most accessible possible condition for an anaphoric subject is: (i) when there is TMA frame continuity from the (adjacent) antecedent's clause, (ii) when the antecedent is 3rd person (for CVC specifically), (iii) when the antecedent is the subject in its clause, (iv) when the antecedent is in an immediately prior clause which is prosodically and syntactically linked to the Intonational Unit containing the anaphor, (v) when the anaphor is in a clause chain-medial position, and (vi) when a short anaphoric distance lies between the anaphor and its antecedent.

The absence of any of these conditions will be probabilistically less favorable to Ø, and SCs will be more likely to occur. An absence of most or all these conditions will promote X2SBJs.

5.3.2 Language-external/extra-linguistic factors.

In this section, I address some of the predictions with respect to potential extra-linguistic and sociocultural constraints. One set of constraints is concerned with non-individual-specific factors, such as the data collection materials and the in-group/out-group identity of the interviewer(s) with respect to the local community. The other set consists of individual-specific language-external constraints related to sociocultural categories associated with each participant.

5.3.2.1 Non-individual-specific language-external constraints.

Among the language-external constraints, some are specific to the individual background of the participant while other constraint do not relate to the participants on an individual basis, but instead relate to the data collection materials or the personnel engaged in data collection. This

subsection deals with latter, namely, the dynamics of the sociolinguistic interview and the picture description narrative as they relate to turn taking, and the in-group or out-group status of interviewer(s) with respect to the local or regional community of the participants.

RQ5: What are the language-external non-individual-specific constraints that influence SPE in CVC? What is the nature of the effect?

H5: The dynamics of the question-and-answer format during a sociolinguistic interview is aimed at eliciting semi-spontaneous speech that at best can only approximate naturalistic conversational style. The "observer's paradox" may be responsible for participants altering their speech style or register, which may impact the distribution of some variable linguistic phenomena (Labov 1970; Wilson 1987).

Furthermore, question-and-answer dynamics have a particular discourse organization and the positioning of an anaphor within a stream of discourse will impact SPE. For example, an anaphor contained in a response to a question during the sociolinguistic interview may favor an overt pronominal realization as opposed to a turn-internal target subject which may favor SCs or Ø. On the other hand, this same effect may not obtain when a speaker voluntarily initiates a discourse 'chunk' during the picture description narrative task, when their speech is unprompted by a question. Observed effects might be related coreferential relationships among subject referents contained in the question and response, or to the discursive/pragmatic effects associated with the question-and-answer format more generally (cf. Gumperz & Berenz 1993).

Wagner (2016) finds a preference for anaphoric Ø turn-initially as the first subject (response) with English 1st person singular subjects. On the other hand, considerations of clause linking and chaining discussed in 5.3.1.2. would indicate that overt pronominals should be favored in turn-initial or clause chain-initial positions (cf. Givón 2017), both contexts associated with question responses.

I posit that in CVC, X2SBJs will be favored in the first clause in a response to a question (in the interview task), and in turn-initial position (not prompted by a question) in the picture description narrative (see Section 5.5.2.4., below). Ø subjects and SCs will tend to occur turn-internally in the interview and picture description task.

Another potential non-individual-specific language-external constraint is the community status/identity of the interviewer. As will be detailed further in the Section 5.5.2.4., for some interviews, I was the sole interviewer. I am a community outsider, not of Cabo-Verdean descent, who has acquired CVC as fourth language. For other interviews, a native speaker community insider was present and participated in asking questions and giving instructions to participants. Given the diglossic dynamics of Cabo-Verdean society, it is conceivable that speakers conversing with a community outsider might tend to employ a more acrolectal register of CVC or engage in code-switching with Portuguese, or introduce certain Portuguese elements into their speech, deliberately or otherwise, as part of an "observer's effect".

I leave pending empirical confirmation, without advancing a prediction, if patterns of SPE are modulate by the identity/community-status/language background of the interviewer(s).

5.3.2.2 Individual-specific language-external constraints.

Among the individual-specific language-external constraints that could have some influence on SPE in CVC are those related to the speakers' age, socioeconomic status, and language background with respect to the regional dialect of CVC in which they were immersed during childhood and adulthood. These issues are closely intertwined with diglossia in Cabo

Verde, which might be characterized as a stratified form of diglossia that resembles the dynamics found in other societies where languages that have traditionally been classified as creoles are spoken, as well as in other African societies where European states imposed top-down linguistic regimes that subjugated and repressed popular linguistic forms in favor of an idealized metropolitan standard. (see Section 2.6).

RQ6: What are the language-external individual-specific factors that influence SPE in CVC? What is the nature of the effect?

H6: There is little prior research to indicate how Socioeconomic Status might modulate SPE in CVC. In Section 2.6, it was noted that CVC exists in a state of diglossia with Cabo-Verdean Portuguese (Rosa 2006; Veiga 2015), but that the post-creole continuum model has been criticized for application to the Cape-Verdean case, particularly when contrasting dialectal varieties of CVC (Baptista 2015).

In contemporary Cabo Verde, there are some disparities in access to Portuguese that are correlated with an individual's access to formal education and educational attainment outcomes across the lifetime. As such, the acquisition of Portuguese among Cabo-Verdeans is directly tied to issues of age, class or Socioeconomic Status, diasporic connections, and rural vs. urban residence, among other factors (Rosa 2006; Coonan 2007).

Since independence in 1975, and then again after the opening of the Republic to multi-party rule in the 1990s, public education has improved substantially in nearly every capacity (Meintel 1984; Reviere 2005). Educational enrollment, literacy, and completion rates are among the highest in Africa for the primary level, though rates lag at the secondary level, as do rates of overall educational attainment (World Bank 2019). As of 2015, the adult literacy rate was 87% overall, but 98% for Cabo-Verdeans between the ages of 15-25 (Education Policy and Data Center 2018). Since education is conducted almost completely in Portuguese, there are likely to be differences in exposure to Portuguese in terms of age and educational attainment. It would not be unreasonable to expect that differences in exposure to Portuguese might have some effect on patterns of linguistic

structural variation in CVC, be they a result of cross-linguistic influence or generalized effects associated with bilingualism. It is unclear, however, in exactly what way patterns of SPE in CVC would be effected. Cabo-Verdean Portuguese is typologically proximate to European Portuguese, though there has been a lack of investigation into patterns of SPE in the Portuguese of CVC-Portuguese adult bilinguals59. If Cabo-Verdean Portuguese does resemble EP in overall NSP status, it would then best be described as a consistent NSL (cf. Duarte 2000; Duarte & Soares da Silva 2016; for SPE in EP).

Would this imply that speakers who receive more input from Portuguese, interact more often in Portuguese, or who have more comprehensively acquired Portuguese, are more prone to analyze CVC SCs as inflectional affixes (equating them to Portuguese person-number inflectional suffixes)? Or as nominative pronominals? Would the use of anaphoric Ø be expected to differ for speakers with greater exposure to Portuguese? What about X2SBJ constructions? These are outstanding empirical questions awaiting testing with little to no prior research to assist in the advancement of specific predictions.

Bouchard (2018) found for Santomean Portuguese that social factors did not appear to have much relevance, the one exception being speakers' education level, which unexpectedly resulted in lower rates of Ø subjects among more educated speakers. We have seen that numerous studies on Brazilian Portuguese have identified an ongoing diachronic process in which this languages continues to move away from consistent NSL status, and has already developed into a partial NSL. This has resulted in increased overall rates of overt pronominals, but also in the loss the NSP cluster of properties (Duarte 1993, 2000, 2004; Modesto 2000; Barbosa, Kato, & Duarte 2005; inter alia). Supporting evidence for this diachronic change in progress was found for Brazilian Portuguese and Puerto Rican Spanish, in both languages younger speakers have increasingly higher rates of overt

59 Costa & Pratas (2015) explore the acquisition of SPE in Portuguese and CVC among three groups: monolingual CVC speaking children (with typical language development), bilingual CVC and Portuguese speaking children (with typical language development), and bilingual CVC and Portuguese speaking children with Specific Language Impairment.

pronouns (Figure 7, Section 4.5.3) (Duarte 1995; Soares da Silva 2011; cited in Duarte & Soares da Silva 2016). In studies on variable SPE in Spanish, mixed results have been found for age in monolingual Latin American varieties (cf. Alfaraz 2016:5-6, and citation therein), including higher rates of overt pronoun expression in speakers under 50 in PRS (Ávila-Jiménez 1995). Changes in 'apparent time' (gradient age differences) in patterns of SPE in CVC might be visible in shifts in the underlying status of SCs (by grammaticalization), which in turn would be indicative of changes in the overall NSL status of the language under approaches like the 'Null Subject Cycle' (Wratil 2011).

To make a general preliminary observation, Cabo-Verdean society shares a lot in common with other societies that suffered the legacy of the trans-Atlantic Slave Trade, post-Slave Trade imperialism, fascist rule in the metropole, Cold War proxy liberation struggles, and finally independence in the last half of the 20th century. The Cabo-Verdean Republic is still young and emergent, and issues from the colonial era and decades of liberation struggle continue to be prominent in contemporary society. This is especially true of the dynamics of language and power in Cabo Verde (see Sections 2.5 and 2.5). Given the language-based power dynamics, differences in educational attainment, and overall exposure to Portuguese by age and possibly Socioeconomic Status, it would be reasonable to expect to observe some effects for SPE in CVC based on these categories for speaker.

5.4 DELIMITING THE VARIABLE CONTEXT AND CODING THE RESPONSE VARIABLE.

An essential component of a variationist sociolinguistic analysis is adherence to the Principle of Accountability. This requires that the researcher define and delimit the variable context within which a speaker's choice of a linguistic variant is manifested (Labov 1983). If one wants to understand which independent predictive factors influence a speaker's selection of a variant for a given linguistic variable of interest, they must isolate the contexts in which that variable is free to vary, but exclude any non-variable contexts.

The variable context is also known as the envelope of variation, and the contexts that comprise the envelope of variation for morphosyntactic variable should share a discursive, functional, and syntactic contextual affinity (near equivalency) (Labov 1972, 1978; Lavandera 1978; Romaine 1980; Winford 1984, 1996; Cheshire 1987; Poplack 1993; Tagliamonte 2006:9-14; Buchstaller 2009; Brook 2018; *inter alia*). Contexts of non-variability (categorical or near-categorical realization) must be excluded on empirical grounds after recording all instance of a variable in a corpus.

Following these criteria, in a study on SPE, the envelope of variation must be limited to those contexts in which two (or more) anaphoric subjects could plausibly be realized, traditionally assumed to be the covert or overt pronominal nominative argument of the verb (the formal mechanism attributed to argument structure presumably having been defined by the researcher *a priori*). The delimitation on the scope of subject pronominal variation is then a matter for empirical testing and verification in any given language, and the more studies that succeed in reproducing a procedure for delimiting the envelope of variation for SPE in a given language, the more certainty one can have that the delimitation is descriptively adequate.

The envelope for a variable morphosyntactic phenomenon such as SPE will differ cross-linguistically. Isolating the nature of the envelope cross-linguistically is among the essential components of Variationist Typology (Torres Cacoullos & Travis 2019). Variationist Typology takes the constraints that characterize the envelope of variation for a given morphosyntactic variable to be the locus of cross-linguistic differences. After having defined and isolated the variable context, the researcher then seeks to identify a set of constraints that probabilistically condition the variable outcome within the envelope, and ascertain whether these constraints can be found to be relevant for the equivalent morphosyntactic variable across languages or language varieties. Constraints found to be predictive for a morphosyntactic variable across languages or dialects might then be hypothesized to form part a speaker's intrinsic knowledge (cf. Carvalho, Orozco & Lapidus Shin 2016:xii).

A primary objective of the present study is to provide the first delimitation of the variable context for SPE in CVC under a variationist sociolinguistic framework. Beyond its application in the current study, this exercise also serves as a template for future investigation into SPE in CVC. In the spirit of the Variationist Typology research program and Claes' (2017) call for a similar, 'Probabilistic Grammar' approach, this procedure should allow for cross-linguistic comparisons with studies that apply this paradigm to explore SPE in CVC, and in other languages and related language varieties.

We have seen that SPE in CVC differs from SPE in languages lacking atonic subject pronominals. This is due to the availability of at least three nominative anaphoric outcomes in SPE (SC, X2SBJ, or Ø). Single tonic pronouns can also occur without a SC (107).

(107) ***Nos*** *ben fazi kuaranta anu di independensia gosi na sinku juliu*
1.PL.WK come make forty year of independence now on five July
'We came up on forty years of independence this past fifth of July.'

However, the primary context for lone tonic pronouns is with individual-level copula *e* 'to be', which is the only predicate in CVC that categorically excludes SCs (example 30, Section 4.1, reproduced here in 108) (Pratas 2004:55-56). Given that SCs do not occur with *e*, subject arguments of this verb were excluded from the envelope of variation, although arguments of *e* were still coded as antecedents for eligible target subjects where appropriate.

(108) a.) *Mi e/era*[60] *spertu*
WK.1S COP clever
'I am/was an clever.'

60 *e* inflects for aspect in CVC; *era* for imperfect, and increasingly *foi* for preterit as well. In my corpus there unexpectedly occurred several instances of SCs with these inflected forms of *e*. Nevertheless, to constrain the scope of the study, all instances of *e* including its inflected forms were excluded from the envelope of variation. This issue, however, is of interest for future investigation, as is SPE with *e* generally.

b.) **N e/era spertu*

CL.1S COP expert

In my corpus, there were only seven of ~8,500 isolated Intonational Units where a singleton tonic pronoun occurred without a SC as the subject of a verb other than *e*. Therefore, the subject realization in which a lone tonic pronoun occurs as a nominative argument of the verb was excluded from the envelope of variation and set aside for future analysis.

Another issue arises with the tripartite classification of overt CVC subject pronouns adopted by Pratas (2004) and Baptista (2002) (Table 21, Section 4.1). In Sections 4.1 and 4.3, I explored the motivations for adopting the tripartite analysis of CVC overt pronominals, which follows the universalist claims of the Typology of Structural Deficiency (Cardinaletti & Starke 1994, 1996, 1999). This cross-linguistic classification identifies morphosyntactic, prosodic, and semantic properties that are predicted to correlate with strong, weak, and clitic forms.

The coding procedure in the current study coded instances of X2SBJs with disyllabic tonic pronouns uniquely from those with monosyllabic ones. Ultimately, however, following the bipartite analysis adopted by Quint (2000a; *inter alia*), I collapsed these variants into a single X2SBJ category. This was done to keep the levels of the outcome (dependent/response) variables manageable for conducting statistical analysis, as increasing the number of outcomes requires correspondingly large increases in sample size in order to preserve statistical power (Pituch & Steven 2005). Further, it was noted the syntactic-distributional claims associated with the pronominal forms under the Typology of Structural Deficiency, like other 'tests' for clitichood, have not been able to resolve the pronoun vs. affix debate for SCs. With respect to the prosodic claims in the Typology of Structural Deficiency, a full examination of these is beyond the scope of the present study and is reserved for future analysis. Therefore, three subject realizations were eligible for inclusion within the envelope, a SC (109), anaphoric Ø (110), or a X2SBJ (111).

(109) ***N****=teni* *dizoitu* *anu,* ***N****=sa* *fazi* *dizinovu na ogostu*
1.SG.CL=have eighteen year 1.SG.CL=TMA make nineteen in august
'I am eighteen years old, I am going to turn nineteen in August.'

(110) *Kriolu*$_i$ *dj=e*$_i$ *ka, ka ta mutu djuda=l na fala,*
kriolu$_i$ TMA=3.SG.CL$_i$ NEG NEG TMA much help=3.SG.CL.ACC in speak
tem ki ta, eeehh.. fala inglês ma.. ***Ø***$_i$ *ta arma djuda sin*
have to TMA HES speak english DM Ø$_i$ TMA gather help yes
'Kriolu does not really help them that much in speech, they have to, uhh… speak English but… [it] does help some, yeah.

(111) *Mas* ***Ami****=****N*** *ta obi mas morna propi*
CONJ 1.SG.STR =1.SG.CL TMA listen more morna INT
'I listen to more morna really.'

X2SBJs where some material intervenes between the tonic pronoun and SC (112) are nonetheless admitted as instances of X2SBJs.

(112) ***El***$_i$ *tipu ki* ***e***$_i$*=sta* *xatiadu ku sapu*
3.SG.WK$_i$ DM 3.SG.CL$_i$=COP angry with frog
'Him, like, he is angry with the frog.'

Another challenge for the researcher attempting to constrain the envelope of variation for SPE in CVC is with respect finiteness. CVC lacks an infinitive morpheme like Portuguese/Spanish verb stem-final *-r*, or any other explicit means of encoding non-finiteness. While the presence of

TMA particles can be taken to be indicative of finiteness, the bare verb can also be finite, with most stative bare verbs receiving a default non-past reading and non-stative bare verbs receiving a past reading (Silva 1990; Quint 2000:235-238; Tavares 2012:16-20). Further, in many embedded contexts verbs remain bare but may nevertheless be finite. This presents a challenge for the analyst of SPE, since there is no single discrete element that can be used to identify finiteness, and finiteness cannot be assumed to be a proxy for where an overt pronominal nominative argument of tensed verb might be expected. Dijkhoff & Mufwene (1989) similarly observed finiteness to be a fleeting or irrelevant category in Papiamentu, among other creole languages.

Indeed, typologists have observed that there is no discrete morpheme type, word category, clausal/hierarchical status, or degree of embedding that is cross-linguistically consistently correlated with finiteness, to the extent that finiteness might be considered an epiphenomenon (cf. Nikolaeva ed. 2007). Formal definitions of finiteness usual reflect the theoretical perspective of the researcher, rather than some quantifiable, discrete, independent linguistic property. In GG frameworks, as Estrada-Fernández (2016:43-44) observes, devices such as nominative case assignment and subject-verb agreement (Cowper 2002), or person, number, and tense marking (Huddleston 1988:44; Hogg 1992:541), are taken as proxies for identifying finiteness.

In their analysis of finiteness in Pima Bajo (Uto-Aztecan), Estrada-Fernández (2016:44-45) adopts Givón's (2001[1984]) 'scalar approach' to finiteness. The scalar approach views finiteness as a non-discrete phenomenon, since languages can rely on more than one method for marking differences in the inter-dependency of clauses. Among these, a language might employ TMA morphology, nominative pronominal case assignment, person-number agreement, nominalization, subordination, and other linking elements, or several of these simultaneously, as correlates of finiteness (see also Nikolaeva 2010). Since under this view finiteness is not conceived of as a discrete but rather a scalar, it applies in differing degrees along a gradient continuum of inter-clausal dependencies (Givón 2017:308-312).

While a full analysis of finiteness in CVC is beyond the scope of the current study, for the purposes of defining the envelope for SPE, the variable context will have to be selected on the

basis of observed variation in the occurrence of a subject pronominal in the subject slot adjacent a verb, independently of the assumed degree of finiteness attributed to that verb. Despite this, there is at least one context that can be excluded on the grounds of nonfiniteness, and several contexts that can be excluded on the ground of the apparent non-argumental (or quasi-argumental) status of their 'subject slot', or their exclusion of one of the possible SPE outcomes in CVC (X2SBJ, SCs, Ø).

One case that is easily excluded does appear to be attributable to nonfiniteness. Several CVC verbs serve as modals or auxiliaries[61], including *bai* 'go', *bem* 'come', *podi* 'be able to', *konsigi* 'manage to', *kre* 'want', *ten ki* 'have to', *gosta* 'like', *sabi* 'know', and *mesti* 'need'. These admit a nonfinite verbal argument and disallow an overt pronominal from intervening. Therefore, all instances of a putative 'subject slot' between modal/auxiliaries and their nonfinite verbal complements are excluded from the envelope. In cases where the speaker used a modal/auxiliary + bare verb, then proceeded to list activities or events using bare verbs with putative target subjects in immediately following IUs, such that they appeared to be instances of ellipsis of the modal/auxiliary, then these were taken to be instance of non-variability and were excluded from the envelope. In (113), first person *N* is attached to complementizer *k(i)*; this anaphoric target was admitted into the envelope as a SC, but the putative subject slot for the verbs *juga* 'play' and *odja* 'watch' are excluded on the ground of ellipsis.

(113) *Asin k=N* *ta* *pudi sai* *ku* *minis asín, juga, odja*
DM COMP=1.SG.CL TMA can go out with guys DM play watch
tilivizôn…
television
'Like that way I can go out with friends like that, play, watch T.V. …'

61 Tavares (2012:37) and Baptista (2002:114) refer these as 'serial verb constructions', but to distinguish from another serial-verb-like construction, I treat these as modal/auxiliary verbs.

A less transparent case is a kind of verbal concatenation structure resembling a serial-verb construction that employs the TMA marker *ta*, which usually serves as the habitual or non-past marker for non-stative verbs, but fulfills several other functions as well. This serial-verb-like construction, to which I will refer as 'incorporated *ta*', consists of a sequence of an initial verb, usually a movement verb like *bai* 'go', a durative verb like *kontinua* 'continue', or locative *fika* 'stay', then *ta*, followed by another verb, yielding a progressive or iterative reading (Tavares 2012:49). It is unclear whether the putative 'subject slot' prior to the second verbal complex (before embedded *ta*) should be taken to be a genuine argumental position (114)-(118).

(114) *E=fika ta djobi sapu*
3SG.CL=stay TMA stare toad
'He remains starring at the toad ~ He remains [and] stares at the toad.'

(115) *Dipôs e=ta fika kunpridu ta seka*
then 3.SG.CL=TMA become elongated TMA dry
'Then it becomes elongated [and] [it] dries'

(116) *El=fika ta anda ta para*
3.SG.CL=stay TMA walk TMA stop
'He remains [and] walks [and] stops.'

(117) *N=ta dizenhá ta diskontra ku nhas muzika*
1.SG.CL=TMA draw TMA relax with my music
'I draw [and] relax with my music.'

(118) *Es es=sta senpri ta promente=n*

3.PL.WK 3.PL.CL.NOM=COP always TMA promise=1.SG.CL.ACC

'They are always promising me.'

While there were some cases where there appeared to be variation between SCs and Ø in the putative medial 'subject slot' of incorporated *ta*-constructions, since X2SBJs never occurred medially in these close serial-like dependencies, the medial position in the incorporated *ta*-constructions was excluded from the envelope and reserved for future analysis (although it was recorded when the putative medial subject of an incorporated *ta*-construction was an antecedent). Future studies exploring 'incorporated *ta*' in CVC might attempt to ascertain the degree to which an overt SC can be inserted in the medial 'subject slot', and if this might be indicative of incomplete or ongoing serialization. Further, the lexical semantics of the auxiliary or initial verb in these sequences may be determinative of the extent of serialization or grammaticalization associated with the 'incorporated *ta*' sequence.

Another site of unusual or unexpected pronominal variation is with embedded *pa*-clauses. Like in Papiamentu (Kouwenberg 1990), CVC complementizer *pa* introduces non-factive embedded clausal complements. However, in CVC the subject position associated with the embedded verb in a *pa*-clause is a site of variation for SPE (119a-b). Even X2SBJs occurred in embedded (119c) and fronted *pa*-clauses (119d,e). Since *pa*-clauses are a relevant site of variation for all three subject realizations in CVC, they were included in the envelope for analysis.

(119) a.) *E$_i$=sata parti pa Ø$_i$ bai si kaminhu*

3SG.CL$_i$=TMA leave COMP Ø$_i$ go his way

'He is leaving to go his own way.'

b.) *Es$_i$* *sa* *tenta bai a-o* *mezmu tenpu p=Es$_i$*

3PL.CL$_i$ TMA try go at-DET same time COMP=3PL.CL$_i$

djobi si=es *ta* *pega-l*

see COMP=3PL.CL TMA catch=3SG.CL.ACC

'They are trying to go at the same time to see if they can catch him.'

c.) *E$_i$=ta* *bai pa* *kel otu* *pa* *Ø$_{i/j}$* *pudi* *serka-l*

3SG.CL$_i$=TMA go along that other COMP Ø$_{i/j}$ be able entrap=3SG.CL.ACC

pa *El$_k$=E$_k$* *tenta bai pa* *li.*

COMP 3SG.WK=3SG.CL try go over here

'He$_i$ goes along that other [way] [for] [him$_i$/them] to be able to entrap it for him$_k$ to try to go over here.'

d.) *El$_i$* *p=E$_i$* *bem* *pega* *sapu, E$_i$* *kai*

3SG.WK COMP=3SG.CL come catch toad 3SG.CL fall

'Him, for him to come catch the from, he fell.'

e.) *Ami$_i$* *pa=N$_i$* *ganha dinheru, N=ta* *laba* *karru*

1.SG.STR$_i$ COMP=1.SG.CL$_i$ earn money 1.SG.CL=TMA wash car

'For me to earn money, I wash cars.'

Just as Kouwenberg (1990) observed for Papiamentu, and as Lipski (1999) observed for several Romance-lexifier Creoles, CVC also makes extensive use of Ø in constructions where the subject takes an arbitrary, generic, or impersonal interpretation. These constructions often employ the TMA marker *ta* and resemble Portuguese and Spanish 'impersonal *se*'. A seemingly non-argumental 'impersonal *ta*' may be constructed with a generic first-person plural or second-person SC or Ø (120), though in many cases there can be an ambiguous interpretation between a genuine argumental and generic reading (121).

(120) *Praia bu sabi e modi, Ø ta trabadja kes kuza asi,*
Praia 2.SG.CL know COP way Ø TMA work those things like that
mas k=U trabadja, Ø ka ta odja dinheru
more COMP=2.SG.CL work Ø NEG TMA see money
'Praia, you know what it's like, [you] work things like that, the more you work [you] don't see [any] money.'

(121) *Xarén go Nu$_i$ ta kotxi ... Nu ta tenti, Ø$_{i/j}$ ta tra farinha*
Grits now 1PL.CL TMA grind 1PL.CL TMA sift Ø$_{i/j}$ TMA remove cornmeal
'Now the grits, we grind them … we sift them, we remove the cornmeal.'

Due to their uncertain argumental status, impersonal *ta* constructions were excluded from the envelope in an attempt to isolate only tokens that undoubtedly bore coreference with a discernable antecedent. As we will see ahead, however, when a nonspecific or indefinite lexical DP served as a clearly discernable antecedent, its coreferential anaphora were admitted to the envelope.

The verb *tem* is used for existential/presentative constructions that are usually marked with Ø (122). However, sometimes generic or ambiguously argumental overt SCs are used (123). These contexts were also excluded due to their non- or quasi-argumental status.

(122) *so ki klima, azvez ten seka*
just COMP climate sometimes PRSNT drought
'(It's) just that the climate, sometimes there is drought.'

(123) *Ø ten txeu kualidadi di vida, na sintidu di bu ten praia, bu ten*
Ø PRSNT a lot quality of life in sense of 2SG.CL PRSNT beach 2SG.CL PRSNT
un bokadinhu di natureza
DET.INDEF little bit of nature
'You have/there is quality of life in the sense that you have/there are beaches, you have/there is a little bit of nature.

One site of cross-linguistic difference that is likely to be determinative for defining the envelope is in the semantic properties associated with overt pronominals. For instance, unlike in most varieties of Spanish[62], but like in Portuguese, the same set of overt subject pronominals that are used for animate, specific, and definite referents, can also be used to refer to inanimate, non-specific, or indefinite ones. When the speaker introduced a clearly identifiable antecedent DP that bore inanimate, nonspecific, or indefinite reference, then the anaphor resuming that antecedent was admitted to the envelope for analysis, and their semantic referential properties were coded as inanimate, nonspecific, or indefinite in accordance with the properties of the antecedent.

In (124) one can observe nonspecific *kriansa* 'children' resumed by Ø, and in (125a) the indefinite DP *otus* 'others' is similarly resumed by Ø. (125a) is also an example of how, in a chain of target subjects, the referent borne by the first antecedent in the chain was determinative for the semantic properties of each target down the chain, until the chain is broken by the introduction of some other referent. (125b) shows that an overt SC can resume an indefinite antecedent. In (126a) an inanimate antecedent is resumed by anaphoric Ø, and in (126b), an inanimate antecedent is resumed by an SC.

62 Toribio (p.c.) reports overt subject pronouns in Dominican Spanish used to resume inanimate antecedents.

(124) *i kriansa$_i$ ta panha, abes Ø$_i$ ta brinka ku=el*

CONJ children$_i$ TMA grab sometimes Ø$_i$ TMA play with=3PL.WK.OBL

'And the kids grab them, sometimes they play with them.'

(125) a.) *Alguns ta ba pa fora, pa interior, Somada, Tarrafal, otus$_i$ ta bai*

Some people TMA go to out to interior S. T. others$_i$ TMA go

praias di mar, Ø$_i$ ta sai ku familia, Ø$_i$ ta bai odja

beach Ø$_i$ TMA go out with family Ø$_i$ TMA go watch

jogu di futibol

game of football

'Some people go out (to the countryside), to the interior, Assomada, Tarrafal, others go to the beach, they go out with their families, they go to watch soccer games.'

b.) *o-ki algên$_i$ ben, es$_i$ ben fla ma Kauberdi e sabi*

when 3.PL.INDEF$_i$ come 3.SG$_i$ come say COMP Cabo Verde COP nice

'When people come they end up saying that Cabo Verde is nice.'

(126) a.) *Morna$_i$ abes ta leba txeu instrumentu, Ø$_i$ ta leba volinhu*

morna$_i$ sometime TMA have a lot instrument Ø$_i$ TMA take guitar

'Morna, sometimes it has a lot of instruments, [it] has guitar.'

b.) *Primeru N=ta tenta midjora-ba kel, kel kistôn di violensa$_i$*

first 3.SG.CL=TMA try improve-TMA that, that question of violence$_i$

li, pamodi e$_i$=ta straga imagen di Kauberdi

here COMP 3.SG.CL$_i$=TMA ruin image of CV

The ability for X2SBJs to resume such semantically referentially deficient antecedents is a question left pending empirical confirmation. This is issues is dealt with in the next Chapter (Section 6.1)

Yet another unusual context is when a lexical DP is topicalized or left-dislocated and separated from the verb by some intervening material (such as pre-verbal perfective TMA marker *dja*, an adverb, or some left-dislocated element). When this occurs, the dislocated lexical DP may or may not be resumed by a SC in the subject slot following the intervening material but immediately prior the verb, its TMA particles, or negation. In cases where the lexical DP is dislocated, separated from the subject slot by some intervening material, but is not resumed by a SC in the subject position, these were coded as an instance of DP + INTERVENING MATERIAL + Ø (127a). These contexts are excluded from the envelope, but as we will see in the next section, were eligible to serve as antecedents to a target subject. The DP + INTERVENING MATERIAL + Ø antecedent is differentiated from instances of when a lexical DP, separated from the subject slot by some intervening material, is resumed by an overt SC in the subject position adjacent the verb or heads within in the VP (these constructions were coded as a X2SBJ antecedent for the factor SURFACE FOR OF THE ANTECEDENT) (127b). It is also differentiated from cases in which a lexical DP occurred immediately adjacent to the verb (in the 'subject slot') or immediately prior to TMA heads within in the VP (and without a SC; these constructions were coded as a LEXICAL DP antecedent for the factor SURFACE FOR OF THE ANTECEDENT).

(127) a.) ***Nha mai***$_{i}$ *na kaza* ***Ø***$_{i}$ *ta fazi txeu ku pexi*
my mom in house Ø$_{i}$ TMA make a lot with fish
'My mom, at home, [she] makes [it] a lot with fish.'

b.) *i* ***katxor***$_{i}$ *també* ***E***$_{i}$ *sta-ba bem prosimadu di kel sapu*
CONJ dog also 3SG.NOM.CL COP-TMA very close to that frog
'and the dog, he was also very close to that frog.'

In this subsection, the variable context has been delimited in such a way as to explore variability between Ø, SCs, and X2SBJs. Whether X2SBJs employed a disyllabic tonic pronoun or a monosyllabic one, these were treated as equivalent and coded into a single category. Other delineations required eliminating contexts where only variation between SCs and Ø (but not X2SBJs) occurs, such as *impersonal ta*, and semi-finite contexts, such as the serial-verb-like *incorporated ta* construction. Another such context is the nominative arguments of copular *e* since this verb only admits Ø or singleton tonic pronouns. Realizations of a singleton tonic pronoun with other verbs were excluded because of their low overall rate of occurrence.

Any target subject eligible for admittance into the envelope is required to have a discernable antecedent that was indexed with some discourse referent; this restriction allows for exclusion of several impersonal or generic contexts and other non-argumental contexts, such as presentative/expletives. However, when an indefinite, inanimate, or nonspecific lexical DP was introduced and then resumed by one of the eligible subject variants (or by several in a coreference chain), these tokens are admitted to the envelope and coded according to the semantic reference of their antecedent. This is crucial because overt pronominals compete with Ø in these contexts, and it is an outstanding empirical question to what degree tonic pronouns or X2SBJs are capable of resuming indefinite, inanimate, and nonspecific antecedents. A final context that was excluded were instances of false starts in which the speaker introduced a subject referent, and sometimes part of a verb as well, but did not complete the clause or utterance, though sometimes the speaker might resume the utterance with a change or modification to the incomplete utterance. False starts were excluded from eligibility as both antecedents and targets, and usually resulted in the exclusion of the next potentially eligible target.

Having delimited the envelope, I turn now to a description of the transcription procedure, as well the process for coding predictor or independent factors (variables) considered as potential constraints on SPE in CVC. The transcription procedure is discussed first because it directly informs the way in which certain predictors were coded.

5.5 TRANSCRIPTION PROCEDURE AND CODING OF THE INDEPENDENT PREDICTOR VARIABLES.

In this subsection, I describe the procedure used for transcribing the speech in the present study. The transcription procedure is described first because the coding of predictor variables related to discourse-organizational constraints, such as linking across International Units and clause-chaining in 'discourse chunks', were directly dependent upon the transcription procedure.

5.5.1 Transcription procedure.

I completed all first drafts of the interview transcriptions (for more on the data collection procedures, see Section 5.6 below). After completing initial drafts, I then reviewed each transcription two to three times, editing and correcting for accuracy. Completed transcriptions were sent to three transcription assistants for further review, editing, and correction. The transcription assistants were informed of the topic under investigation, trained on the relevant conventions and procedures, and shown several examples of completed transcriptions and their accompanying recordings.

Up to the first 15 minutes of the interview portion of each recording were transcribed. Participants also completed a picture-description narrative task (see Section 5.5.2.4., below) consisting of the 'The Frog Story' (Mayer 1967). 'The Frog Story' is a series of unannotated illustrations presented in a narrative progression in which a boy and his dog go to a pond to catch a frog. The boy, the dog, and the frog are the only three characters, and participants were invited to verbally narrate the progression of events in the story. 'The Frog Story' has been used in a range of linguistic and psychological studies (cf. Sánchez 2003; Slobin 2004; *inter alia*). For interviews collected from participants who also completed the picture description narrative task, the entirety of the 'Frog Story' was also transcribed.

Each completed transcription was edited and corrected by one of three assistants. Two of the transcription assistants are native speakers of CVC who were born and raised in Picos. Another is a near-native second language learner of CVC who has worked and resided in Praia for more

than 15 years; this assistant is a native speaker of US-English and speaks Portuguese at an advanced level as an L3. Of the two native speaking consultants, one was also the same colleague that assisted in identifying, recruiting, and interviewing participants in the second summer. Once the transcription assistants returned the edited and corrected transcriptions, I reviewed them once more and adjusted them to reflect a 'prosodic transcription' format.

The prosodic transcription procedure followed Chafe (1993, 1994), Du Bois *et al.* (1993), Gumperz & Berenz (1993), and Torres-Cacoullos & Travis (2019). The objective is to represent the intonational and discursive properties of conversational speech and/or the question-and-answer format of a traditional sociolinguistic interview. Underlying prosodic transcription is the notion that the concept of the sentence is not independently supported by quantifiable linguistic properties in naturalistic speech (cf. Chafe 1994). Instead, each line of text on the transcription document represents an Intonational Unit (IU), a steady stream of speech contained within a single intonational contour (see examples 95-99, Section 4.5.3).

A rising or steady intonation at the end of an IU is indicative of continuing intonation, while falling intonation at the end of an IU is indicative of final intonation. Continuing intonation was represented orthographically by a comma (,) while final intonation was represented orthographically by a period (.). When an IU follows a previous IU that has continuing intonation, these IUs are taken to be prosodically linked. However, linking only obtained when there was coreference across antecedent and anaphor in adjacent IUs. Torres Cacoullos & Travis (2019) found that 'prosodic linking' promoted the realization of English Ø in IU-initial position. 'Syntactic linking' referred to the use of a coordinating conjunction at beginning or end of an IU that connected adjacent clauses. Torres Cacoullos & Travis (2019) also found that adjacent IUs with a coreferential antecedent and anaphor that were both prosodically and syntactically linked, were the most favorable condition for the English Ø subjects.

For the present study, CVC coordinating conjunctions *i* 'and', *ma(s)* 'but', *(ã)ntôn* 'so, then', *dipôs* 'then', discourse markers *gó* 'now/well/then' and *tipu (ki)* 'like, sort of like', and complementizer *pamo(di)* 'because' were coded as syntactic linkers. In order for an IU containing

an antecedent to be linked in any way to an IU containing the anaphor, they had to be adjacent. If an anaphor in an embedded clause had an antecedent in the matrix clause, these were always coded as both syntactically and prosodically linked (unless they fell in distinct but adjacent IUs separated by a pause, in which case they were coded as only syntactically linked).

In (128)-(131), there are examples of adjacent IUs with coreferential antecedents and anaphora that are prosodically linked (128a-c), syntactically linked (129a-c), both prosodically and syntactically linked (130a-c), and unlinked (131a-c), each with a Ø, SC, and X2SBJ realization. Many IUs contain at least one full clause, but can contain more than one, and do so often in cases of syntactic embedding and coordination.

(128) Prosodic linking

a.) ***E***$_{i}$*=ta* *pila kana**,***

3.SG.CL =TMA press cane

Ø$_{i}$ *ta fazi si groginhu*

Ø$_{i}$ TMA make his sugarcane liquor

'He presses the sugar cane, [he] makes his sugarcane liquor.'

b.) ***N***$_{i}$*=ta* *ferbi nha kafé pretu**,***

1.SG.CL$_{i}$=TMA boil my coffee black

N$_{i}$*=ta* *bebi*

1.SG.CL$_{i}$=TMA drink

'I boil my black coffee, I drink [it].'

c.) ***Es***$_{i}$***=Es***$_{i}$ *sta la**,***

3.PL.WK=3.PL.CL COP there

Es$_{i}$***=es***$_{i}$ *ten ki ntegra na sosiedadi*

3.PL.WK=3.PL.CL have to integrate in society

'The are there, they have to integrate into the society.'

(129) Syntactic linking

a.) *iii* *dj=**E***$_{i}$ *ta* *sigi* *kel* *pasu**.***

CONJ TMA=3.SG.CL TMA follow those footprints

i ***Ø***$_{i}$ *odja rãn sta d-un* *ladu*

CONJ Ø$_{i}$ see frog COP on-DET.INDEF side

'aaand now he follows those footprints. And [he] sees that frog is [there] to one side.'

b.) ***Rapzinhu***$_{i}$ *buâ* *n-el**.***

boy jump on-3.SG.WK.OBL

i *li* ***E***$_{i}$*=sa* *ba ku* *redi p=E*$_{i}$ *panha=l*

CONJ here 3.SG.CL=TMA go with net COMP=3.SG.CL catch=3.SG.CL.ACC

'The boy jumped on him. And here he is going with the net to catch it.'

c.) *Ali* *e* *abitasôn* *di* ***sapu***$_{i}$***.***

here COP inhabitance of frog$_{i}$

i ***El***$_{i}$ *dj=**e***$_{i}$ *fika* *tristi, ne?*

CONJ 3.SG.WK$_{i}$ TMA=3.SG.CL$_{i}$ become sad DM

'Here is the frog's place of inhabitance. And here he becomes sad, right?'

(130) Both

a.) ***Minis di*** ***Praia***$_{i}$ *e* *so* *si**,***

guys from Praia COP only like that

ma**=**Es$_{i}$ *ta* *arma bon karaka* *pa*

but=3.PL.CL TMA put on good beachside party DM

'Guys from Praia are just like that, but they put on a good beachside party man.'

b.) ***Kel sapu***$_{i}$ *ta fudji mas un bes,*

that frog$_{i}$ TMA escape more DET.INDEF time

i Ø$_{i}$ *ba fika sukundidu riba d-un pedra*

CONJ Ø$_{i}$ go stay hidden on top of-DET.INDEF rock

'That frog escapes one more time, and [he] goes to hide on top of a rock.'

c.) ***Es***$_{i}$ *ka sa da txeu importansia pa violensa ki sta na sosiedadi,*

3.PL.CL.NOM NEG TMA give a lot importance to violence that COP in society

pamodi Aes$_{i}$***=Es***$_{i}$ *ta anda ku ses siguransa na karru*

COMP 3.PL.STR$_{i}$=3.PL.CL$_{i}$ TMA go with their security in car

'They aren't putting a lot of importance on the violence in society, because they go around safe in their cars.'

(131) No linking

a.) ***N***$_{i}$*=sta pa pega autukarru número tres.*

1.SG.CL=intend to.PROG catch bus number three

kuaranta minutu ***N***$_{i}$*=ta txiga na bo*

forty minute 1.SG.CL=TMA arrive on 2.SG.WK.OBL

'I am planning to catch bus number 3. Forty minutes, I'll arrive where you are.'

b.) *Kel otu anu* ***M***$_{i}$*=ba trabadja MDR* **…**

that other year 1.SG.CL=go work MDR

Ø$_{i}$ *ba lora pedra, koba txôn, dipôs...*

Ø$_{i}$ go mine rock dig earth then

'That next year I went to work at MDR… [I] went to mine rocks, dig up earth, then…'

c.) *Ulisis, ki* ***El*** *e prizidenti di kâmara, li di Praia.*

U. COMP 3.SG.WK COP president of chamber here from Praia

El$_i$=***e***$_i$ *kumesa fazi un bon trabadju,*

3.SG.WK=3.SG.CL begin do DET.INDEF good job

kalseta rua, fazi prasas, fitines park, asi.

cobble street make plaza fitness park like that

'U., who is the leader of parliament, here from Praia. He started to do a good job, cobbling the streets, making plazas, fitness parks, things like that.'

The interviewers' questions were also included in the transcription. This allowed the interviewers' speech to be included in measurements of anaphoric distance and in the tracking of antecedent-anaphor relationships for the purpose of recording the antecedent accessibility pattern and priming effects for target subjects in response turn-initial position. However, anaphora in clauses contained in a response to a question were always coded as unlinked. Turn-initial participant-initiated discourse chunks were necessarily coded as never prosodically linked, even in the event of coreference with an antecedent contained in a prior discourse chunk.

5.5.2 Coding the predictors.

The independent predictors/factors/variables are organized into those related to the discourse, referent tracking, clausal configurations, coreference and binding relationships 5.5.2.1.; the morphosyntactic status and referential semantics of antecedents 5.5.2.2.; the verb and verb phrase 5.5.2.3.; and language-external factors 5.5.2.4. A table summarizing the relevant predictors accompanies the descriptions for the variables at the end of each subsection.

5.5.2.1 Predictors related to discourse organization, coreference, referent tracking, the clause, and relationships between clauses.

Overt and covert pronominals allow for the tracking of referents across the discourse and are therefore constrained by the online cognitive processes associated with this task in everyday verbal exchanges. As a result, the realization of a subject form must be sensitive to domain-general cognitive constraints that affect linguistic and non-linguistic tasks alike, such as priming (persistence) and working memory (as it relates to referent activation, i.e. anaphoric distance), but also to language-specific constraints on referent tracking like coreference and binding relationships, switched vs. continued reference, overall antecedent accessibility, the clausal configurations in which a target its antecedent are contained, and the syntactic role an antecedent fulfills in its containing clause.

The predictor factor ANTECEDENT ACCESSIBILITY PATTERN referred to the syntactic role of the antecedent in its clause, as well as the clausal configuration that lay between the target and its antecedent. This category followed Duarte's & Soares da Silva's (2016:7-8) classification with some modifications (Table 29).

Table 29. The levels of the factor ANTECEDENT ACCESSIBILITY PATTERN

PATTERN	DESCRIPTION
PATTERN A	The anaphor is in an embedded clause and is c-commanded by its antecedent (132)
PATTERN B	The antecedent is the subject of the immediately prior clause (133)
PATTERN C	The antecedent is in the prior clause, but has some role other than the subject (134)
PATTERN D	The antecedent is the subject of a clause that is separated from the clause containing the anaphor by at least one clause, and the intervening clause contains a verb whose subject bears a referent other than that of the target anaphor and its antecedent (like PATTERN B but in a nonadjacent clause) (135)
PATTERN E	The antecedent plays a role in the clause other than the subject and is separated from the clause containing the target anaphor by at least one intervening clause containing a subject that bears some referent other than that of the anaphor and its antecedent (like PATTERN C but in a nonadjacent clause) (136)
PATTERN F	The anaphor is in a main clause following a fronted adverbial subordinate, *if*-clause, or *pa*-clause, containing the antecedent (137)
PATTERN G	The antecedent is in a prior, separate, and distinct clause-chain or 'discourse chunk' (the referent is an old discourse topic) (138)

(132) Pattern A

maioria di berdianus$_{i}$*, na fin di simana,* ***es***$_{i}$*=ta bati* ***pamodi*** *dia di simana* ***es***$_{i}$*=ta trabadja*

‘Most Cabo-Verdeans on the week they wash clothes because during the week they work.’

(133) Pattern B

pesoas$_{i}$ *sata fitxa txeu odjus pa kes kuza la,*

Ø$_{i}$ *sata finji ma Ø*$_{i}$ *ka sata odja*

‘People close their eyes to those thing there,

[they] pretend that [they] aren’t seeing [it]’.

(134) Pattern C

*na fornu dja=nu ta pui-****Ø***$_{i}$ *keju,*

Ø$_{i}$ *ta fika mas sabi propi*

‘In the oven now we put more cheese [on it],

[it] ends up tasting even better.’

(135) Pattern D

um bes ***N***$_{i}$*=ta jugaba oril.*

mas gosi li tipu k-e un prâktika ki kuazi sta dizaparesi,

asi dipôs també falta di tenpu.

mas ***N***$_{i}$ *gosta txeu Oril.*

'Once I used to play Oril.

But now it's like it's a practice that is almost disappearing,

like then there's also a lack of time.

But I like Oril a lot.'

(136) Pattern E

podi ser munti grandi prigu un kriansa$_{i}$ *sozinhu pamodi* ***el***$_{i}$ *so ku* ***si***$_{i}$ *kãu,*

kãu$_{j}$ *ka pudi ten kel maturidadi así kel intelijensia,*

si ***Ø***$_{i}$ *fogaba lisin pa Ø*$_{j}$ *djudaba el*$_{i}$

'[It] can be very dangerous, a child by himself/herself, just him/her and his/her dog,

[the] dog might not have the maturity, like, the intelligence,

if [he/she] were to drown here, for [it] to help him/her.'

(137) Pattern F

oras ki ***Mi***$_{i}$*=****N***$_{i}$ *ta le=l,*

N$_{i}$*=ta atxa=l ma un istória típiku*

'when I read it, I find it's a typical story.'

(138) Pattern G

li nu ten un imajen, podi ser kasador,

e=teni un kãu,

i li sa parsi m-e, m-e pexi ki ***es***$_{i}$ ***es***$_{i}$ *tra=s.*

ali tem un lagoua, ki ten un sapu, e n-um boski.

... -TURNS PAGE- …

i li ***es***$_{i}$*=ta korri tras di sapu*

'Here we have an image, [it] could be a hunter,

he has a dog,

and here it seems that [it] is, that [it] is fish that they got (them)

here we have a pond, that has a frog, it is in a forest.

... -TURNS PAGE- …

And here they run up behind the frog.'

Antecedent accessibility patterns have been shown to be highly determinative for SPE in Portuguese (Barbosa, Duarte, & Kato 2005, *inter alia*), English (Wagner 2016), and Spanish (Martínez Sanz 2011, and references therein; Otheguy & Zentella 2012, and references therein; Carvalho, Orozco, & Lapidus Shin eds. 2015, and references therein). Switch-reference (PATTERNs D, E, and G), along with contexts involving non-subject antecedents (PATTERNs C and E), are likely to favor overt SPE over anaphoric Ø, and in particular to favor X2SBJs over SCs and Ø. The antecedent accessibility pattern most favorable to Ø has been shown to differ across Romance languages; configurations involving a c-commanding antecedent (PATTERN A) were considered a "resistance context" that still promotes the expression of Ø in partial NSLs like Brazilian Portuguese and Puerto Rican Spanish (Duarte & Soares da Silva 2016:22).

The predictor SURFACE FORM OF THE ANTECEDENT encoded the morphological category (part of speech) of the antecedent, regardless of its syntactic position. This was included to account

for the effects of structural priming. Linguistic structural priming has been repeatedly shown to exert a powerful psychological effect on the repetition of linguistic forms (Bock 1986; Logbell & Bock 2003; *inter alia*), not just in the laboratory setting, but also in everyday conversational discourse (Weiner & Labov 1983), as a catalyst for cross-linguistic influence in contexts of language contact (Poplack 1980), in corpora (Szmrecsanyi 2005), and has been shown to be an active constraint for SPE in Spanish (Cameron 1993; Flores-Ferrán 2002; Cameron & Flores-Ferrán 2004; Travis 2007; *inter alia*), Portuguese (Bouchard 2018), English (Wagner 2016; and references therein), and in these languages in contact with each other (Carvalho & Child 2011; Carvalho & Bessett 2015; Abreu 2012; Torres Cacoullos & Travis 2015). Priming is likely to be an active constraint for SPE in CVC as well.

The levels for this predictor were LEXICAL DP, DOUBLE SUBJECT, CLITIC, TONIC pronoun (strong or weak), anaphoric Ø, the LEXICAL DP + INTERVENING MATERIAL + Ø construction (see example 127, Section 5.4), the putative medial subject position of the INCORPORATED *TA* construction[63] (see examples 114-118, Section 5.4), a POSSESSIVE PRONOUN marking person-number concord with the possessor in in a lexical DP (e.g. ***nha*** *mai* 'my mother', ***si*** *katxor* 'his dog', etc.), or one of the oblique tonic pronouns that serve as the argument of preposition *di* in possessive constructions involving lexical DPs, as in DP + *di* ***meu*** 'of mine', *di* ***bo*** 'of you' (your), *di* ***sel*** 'of him/her' (his/her), *di* ***nos*** 'of us' (our), and *di* ***ses*** 'of them' (their). Most speakers also used a lexically specific construction, apparently a grammaticalized collocation, that repeatedly appeared with superstrate inflection in the first the person, *ax****u*** (<*acho* Pr.) 'I think'. This collocation always had Ø in the subject slot and occurred with complementizer *ki* (contrast this with the equivalent CVC construction *N ta atxa ma* 'I think that', with the uninflected verb stem, TMA particle *ta,* and complementizer *ma*). While the subject argument of *axu* was excluded from the envelope, the final /u/ was taken to be an inflectional suffix bearing 1st person referential

63 This was done to account for the unclear status of the 'subject slot' in the middle of the incorporated *ta* construction (see Section 5.4).

features and was thus eligible to serve as an antecedent to a target subject[64]. This category was represented by the level INFLECTION.

The predictor factor prosodic/syntactic LINKING was included since it is associated with maximal referential continuity and therefore promotes the expression of Ø and atonic pronominals (Givón 2001[1984]:417-419, 2017:7-8). Additionally, we saw in Torres Cacoullos & Travis (2019) that prosodically and syntactically linked adjacent IUs containing a coreferential anaphor-antecedent promote Ø subjects in English. Outcomes for this variable were NO LINK, PROSODIC LINKING, SYNTACTIC LINKING, and BOTH (prosodic and syntactic linking).

Linking could only apply when an IU containing a target was adjacent the immediately prior IU containing its antecedent. Prosodic linking was associated with steady or rising intonation in at the end of the IU containing the antecedent and represented by a comma in the transcriptions. Syntactic linking involved coordinating conjunctions *i* 'and', *ma(s)* 'but', *(ã)ntôn* 'so, then', *dipôs* 'then', discourse markers *gó* 'now/well/then' and *tipu (ki)* 'like, sort of like', and complementizer *pamo(di)* 'because' were coded as syntactic linkers (see examples 128-131, Section 5.5.1, above). Embedded nominative anaphora whose antecedent was in the matrix clause were always coded as BOTH, except in the rare case that the speaker made an audible and sustained pause between the matrix and embedded clauses, initiating the embedded clause within a separate and distinct (but adjacent) IU; in that case, the clauses would be coded for SYNTACTIC LINKING only.

The factor ANAPHORIC DISTANCE describes the distance in number of words between a target subject and its antecedent. ANAPHORIC DISTANCE has been shown to condition SPE cross-linguistically (Givón 2001[1984]:463-464, 2017:13-18; for Spanish, Otheguy & Zentella 2012; and references therein; for Portuguese, Duarte & Soares da Silva 2016; and references therein). Short anaphoric distances are expected to favor Ø subjects. While anaphoric distance is in a sense already represented by the non-adjacent clausal configurations in antecedent accessibility patterns like D and E, measuring anaphoric distance in number of words has proven to be a predictive and

64 See Givón (2001, 2017) and Siewierksa (2004) for the pronominal status of bound inflectional affixes, at least with respect to referent tracking.

more accurate indicator of morphosyntactic variation (Szmrecsanyi 2005; Hinrichs, Szmrecsanyi, & Bohmann 2015). As such, ANAPHORIC DISTANCE was included as a continuous predictor variable measured in number of words between an anaphor and its antecedent.

CLAUSE CHAIN POSITION encoded the position of the anaphor's clause within a 'discourse chunk' (a series of clauses unified under a common thematic and prosodic contour), relative to the beginning, middle, and end of the discourse chunk. Givón (2001[1984], 2017) provides robust cross-linguistic evidence that the position of a subject in a clause chain is relevant for determining the outcome of SPE. In languages with atonic subject pronouns and Ø, these forms tend to appear in positions of maximal referential continuity, such as a clause-medial position in a clause chain, rather than in chain-initial, chain-final, or "discourse grounding" clauses (Givón 2001[1984]:434).

There could possibly be redundancies between the factors CLAUSE CHAIN POSITION and ANTECEDENT ACCESSIBILITY PATTERN, but since Ø is associated with even higher referential continuity than atonic pronouns, Ø will be preferred over SCs in chain-medial positions where there is also referential continuity (no switch reference). CLAUSE CHAIN POSITION differs from LINKING in that the effect of CLAUSE CHAIN POSITION on the anaphor is not contingent on coreference (linking can only apply to coreferential antecedent-anaphor in adjacent IUs). The levels of the predictor CLAUSE CHAIN POSITION were GROUNDING CLAUSE, CHAIN-INITIAL, CHAIN-MEDIAL, and CHAIN-FINAL. GROUNDING CLAUSES often involve adverbials, existential, and presentative constructions and serve as "coherence bridges" between distinct discourse chunks and competing themes across the discourse (Givón 2001[1984]:356).

CLAUSE TYPE may be a highly relevant predictor of SPE in CVC. Clausal embedding can be associated with reduced finiteness, and "clause-rank discontinuity" in the form of subordinate-main clause subject switching, which often promotes the use of overt pronouns (cf. Givón 2001:383-386). However, languages differ greatly in how they treat SPE in embedded contexts (Givón 2001[1984]:39-89; see Duarte & Soares da Silva 2016 for NSLs in Romance). Further, we saw contradictory evidence with respect to where Ø should be most productive in CVC; Costa & Pratas (2008, 2013) assert that Ø is excluded from root contexts and that anaphoric Ø is only

possible in embedded contexts when bound by an indefinite or WH-operator. Lipski (1999) observes Ø almost exclusively in main clauses for several Iberian-lexifier creoles, though he argues that these instances of Ø are referentially ambiguous, at best quasi-referential, or displaying fully generic or impersonal readings. This leads him to associate the formal category underlying Ø in these languages with arbitrary *pro*. Kouwenberg (1990) attributes matrix Ø in Papiamento a similar arbitrary or quasi-referential status, though she finds anaphoric Ø to be possible in some embedded *pa*-clauses. However, all the examples that Baptista (2002) gave of anaphoric Ø in CVC are in matrix contexts. To test these contradictory claims, CLAUSE TYPE was coded according to the levels MAIN, SUBORDINATE (with complementizers *ki* and *ma*), *SI*-CLAUSE (if-clause with complementizer *si*), ADVERBIAL SUBORDINATE, RELATIVE, *PA*-CLAUSE, COPULAR COMPLEMENT, QUESTION, and COORDINATE.

Table 30. Predictors related to discourse organization, coreference, referent tracking, the clause and inter-clausal relationships.

Predictor	Levels
ANTECEDENT ACCESSIBILITY PATTERN	PATTERN A, PATTERN B, PATTERN C, PATTERN D, PATTERN E, PATTERN F, PATTERN G
SURFACE FORM OF THE ANTECEDENT	LEXICAL DP, TONIC, CLITIC, Ø, DP + INTERVENING MATERIAL + Ø, X2SBJ, *DI* OBLIQUE[65], POSSESSIVE PRONOUN, INFLECTION, INCORPORATED *TA* MEDIAL SUBJECT
LINKING	PROSODIC, SYNTACTIC, BOTH, NO LINK
CLAUSE CHAIN POSITION	GROUNDING, INITIAL, MEDIAL, FINAL
CLAUSE TYPE	MAIN, SUBORDINATE, *PA*-SUBORDINATE, *SI*-CLAUSE, ADVERBIAL SUBORDINATE, RELATIVE, QUESTION, COPULAR COMPLEMENT, COORDINATE
ANAPHORIC DISTANCE	continuous numeric

[65] This level is eventually merged into the TONIC level following Analysis 1 (see Sections 5.7.; 6.1.; 6.2.).

Having enumerated the independent predictors related to issues of discourse organization, referent tracking, coreference, clause type and inter-clausal relationships, I turn now to those factors that are related to the morphological composition and semantic referential properties of the antecedent.

5.5.2.2 Predictors related to the morphological composition and semantic-referential properties of the antecedent.

All target anaphora were coded for PERSON/NUMBER. 2nd person referents occurred at low overall rates in the corpus and with little variation when directed at a specific interlocutor[66] (the interviewers) and were thus excluded from the envelope. The levels for PERSON/NUMBER were 1SG, 1PL, 3SG, and 3PL.

In Brazilian Portuguese, Duarte & Soares da Silva (2016:15-16) found the highest rates of Ø with 3rd person referents, and the highest rates of overt subjects in the 2nd person; for Puerto Rican Spanish, the highest rates of Ø were with informal second person *tú*, although overall rates of Ø for all person-number instantiations were within 15% of the rate for 2nd person realizations. In many other languages with anaphoric Ø, but which do not identify Ø by way of person-number inflection, 1st and 2nd person pronouns tend to be dropped more often, the reasoning being that, as SAP pronouns, they are highly topical and thus highly accessible for resumption by anaphoric Ø (cf. Wratil 2011). Baptsita (2002:257-259), in discussing CVC specifically, provided examples of Ø in all person-number configurations. She also argues that "default third person features in AGR" allow for identification of Ø in absence of SCs or overt tonic pronominals (2002:258).

Each target anaphor was also coded according to its antecedent's ANIMACY (or the animacy of an initial antecedent in a coreference chain). Any subject with a human or animal referent was coded as ANIMATE, while subjects that referred to nation-states, governments, corporations, or

66 There was observable Ø/overt variation for generic and arbitrary second person referents, but these have been excluded from the envelope; they are reserved for future analysis.

other bureaucratic entities were coded as COLLECTIVE. All other referents were coded as INANIMATE.

Animacy is known to be relevant for SPE cross-linguistically. Since SAP pronouns are always animate, the 1st person will tend to favor Ø in languages that rely strictly on discourse antecedents for the identification of anaphoric Ø (cf. Wratil 2011, for cross-linguistic evidence; Wagner 2016, for English). However, inanimate reference has been found to promote Ø in both European and Brazilian Portuguese (Kato & Duarte 2003, 2005; Duarte & Soares da Silva 2016). Under the Typology of Structural Deficiency, *pro* is considered a deficient pronoun and thus admits inanimate reference, as do clitics. Kato's & Duarte's (2003, 2005) "avoid pronoun referentially deficient pronouns principle", if found to be active in CVC, would promote anaphoric Ø after inanimate antecedents.

The referential value of subject is also tied to issues of specificity and definiteness. These two notions were coded into a single category with levels +DEFINITE +SPECIFIC (e.g. *nha mai* 'my mom'), +DEFINITE -SPECIFIC (e.g. *jovens* 'the youth'), -DEFINITE +SPECIFIC (*um armũn di meu* 'a brother of mine' ~ 'one of my brothers'), and -DEFINITE -SPECIFIC (*um gadju* 'some guy'). In all cases of nonspecificity or indefiniteness (as with inanimacy), it was the semantic value of the antecedent that determined the coding outcome for the anaphor, and in coreference chains, it was the initial antecedent in the chain that determined the referential specifications for each target down the chain until the chain was broken by the introduction of some alternative discourse referent (example 125a, Section 5.4). Just as with animacy, under the Typology of Structural Deficiency, strong forms (fully referentially specified) are expected to be disassociated from nonspecific and indefinite reference, as opposed to deficient (weak and clitic) forms, which are capable of bearing nonspecific and indefinite reference, since they receive specification from an antecedent (Cardinaletti & Starke 1999). Duarte & Soares da Silva (2016:18) found that nonspecific reference promoted Ø in BP.

Table 31. Predictors related to the referential properties of the antecedent

Predictor	Levels
PERSON/NUMBER	1SG, 1PL, 3SG, 3PL
ANIMACY	ANIMATE, COLLECTIVE, INANIMATE
SPECIFICITY/DEFINITENESS	[+SPECIFIC, +DEFINITE], [+SPECIFIC, -DEFINITE], [-SPECIFIC, +DEFINITE], [-SPECIFIC, -DEFINITE]

Having addressed the predictors associated with the antecedent morphological composition and semantic-referential properties, I turn now to description of the predictor factors related to the verb and the verb phrase.

5.5.2.3 Predictors related to the verb and the verbal domain.

The remaining language internal variables were all related to the verb phrase. The constraint TMA referred to the surface TMA marking of the verb for which the target subject was an argument and included the levels BARE, *TA*, *-BA*, *SATA*[67], *DJA*, INCORPORATED *TA*[68], IRREGULAR, and COMBINATION. The level IRREGULAR referred to verbs that receive superstrate-like inflection for aspect, in some cases with both imperfective and preterit forms, in others just imperfective or just preterit; among these are *tinha/tivi* 'had', *sabia* 'knew', *podia/pudi* 'was able to/did', *kria/kix* 'wanted', *stivi* 'was', and *binha* 'came'. The level COMBINATION referred to any combination of more than one TMA particle, an auxiliary/modal and a TMA particle, or irregular forms with a TMA particle and/or auxiliary/modal, or any of the above occurring simultaneously.

The variable VERB CLASS was based on the classification in Silva (1990), which itself is a modification of the stativity-based categorization in Bickerton (1975:128). This classification

67 The progressive marker could also be realized as *sta, sa,* or *ata.*

68 This referred the cases where the subject was the argument of the auxiliary/modal or initial verb in the incorporated *ta* sequence, not the putative subject slot of the incorporated or 'main' verb.

builds on the observation that in many languages historically classified as creoles, "the zero form [of the verb] marks the past of action verbs and nonpast for state verbs." To this distinction, for CVC, Silva (1990:146-147) adds the criteria of whether the verb is imperative or controllable, and if the verb can occur with the progressive marker *sta*[69].

CLASS 1 consists of fully stative verbs, they get a default nonpast reading when bare, are not imperative/controllable, and do not admit *sta*. CLASS 2 consists of stative verbs that do not take a past reading when bare and cannot be imperative/controllable, but do admit *sta*. CLASS 3 also consists of stative verbs in that they cannot be imperative/controllable[70], but they admit *sta* and take a default past reading when bare. Finally, CLASS 4 verbs are dynamic/action verbs, they can be imperative/controllable, take a default past reading when bare, and occur productively with *sta*.

In the present corpus, instances of Ø were admitted into the envelope if they clearly occurred as the argument of a finite verb. Given that CVC does not explicitly encode finiteness, this included bare verbs from CLASSES 1 (139) and 2 (140) that get a nonpast reading, and bare verbs from CLASSES 3 (141) and 4 (142), that get a simple past reading.

(139) *ma E$_i$ kai ... **Ø**$_i$ sta dentu di mar*
DM 3SG.CL$_i$ fall Ø$_i$ COP.NPST in sea
'and He fell ... He is in the sea.'

(140) *Ami e gemiu k-um femia$_i$, ia, **Ø**$_i$ txoma Nilda*
1SG.STR COP twin with-DET.INDF female DM Ø$_i$ call.NPST N.
'I am twins with a woman, yeah, [her] name is Nilda'

69 See Tavares (2012:16-22) and Baptista (2002:76-98) for critiques of this classification, and Quint (2000a:248-252) for an alternative classification as "strong" and "weak" verbs.

70 Silva (1990:148) concedes that several Class 3 verbs can be imperative/controllable in negative polarity contexts.

(141) *E=tivi si primeru fidju k=E$_i$ mori na tropa,*

3SG.CL$_i$=have his first child COMP=3SG.CL$_i$ die.PST in army

Ø$_i$ mori ku 18 anu

Ø$_i$ die.PST with 18 year

'He had his first child who died in the army, [he/she] died at 18 years old.'

(142) *Rapazinhu sta xinta, **Ø** ba toma si banhu*

boy TMA sit Ø$_i$ go.PST take his bath

'The is sitting, [he] went to take his bath.'

However, bare verbs from CLASSES 3 and 4 that could not be attributed a past reading (143) were excluded from the envelope[71], as these might be considered not fully finite, referential, or argumental.

(143) *o-ki txiga oras di bai skola, N$_i$=ta subi kaza,*

COMP arrive time to go school 1SG.CL=TMA go up house

Ø$_i$ txiga, Ø$_i$ laba, Ø$_i$ almusa

Ø arrive.NPST Ø wash.NPST Ø lunch.NPST

'Once it's time to go school, I go up to my house, I arrive, I wash, I lunch.'

There is no one-to-one correlation of TMA marking and the actual TMA interpretation attributed by a speaker/hearer to a given proposition in CVC (because of the behavior of bare verbs like in examples in 139-143, but also because several TMA markers exercise more than one function). Therefore, the predictor TMA FRAME sought capture the actual TMA interpretation and

[71] As we will see in the section on statistical procedures, instances of Ø anaphora like in (143) were removed following the preliminary exploratory analysis (Analysis 1) and were excluded from all subsequent anlyses (2-4) that formed the basis for the conclusions ultimately drawn in the final chapter.

disentangle this from the surface TMA marking. The levels for TMA FRAME included NONPAST, PRESENT PROGRESSIVE, SIMPLE PAST, IMPERFECTIVE, PERFECTIVE, FUTURE, and IRREALIS. The levels PAST PROGRESSIVE and HABITUAL PROGRESSIVE were recorded, but due to the low number of observations, the former was merged with IMPERFECTIVE and the latter with PROGRESSIVE.

The predictor TMA FRAME SWITCH sought to account for any reduction in antecedent accessibility incurred when there is a switch in TMA frame between an antecedent's clause and its target anaphor's clause. This predictor included the levels YES, NO, and X; this latter level was for instances in which the antecedent was not a nominative argument of the verb in its containing clause.

Table 32. Predictors related to verb and the verb phrase

Predictor	Levels
TMA	BARE, AUX/MODAL, *TA*, *-BA*, *SATA*, *DJA*, INCORPORATED *TA*, IRREGULAR, COMBINATION
TMA FRAME	NONPAST, PRESENT PROGRESSIVE, SIMPLE PAST, IMPERFECTIVE, PERFECTIVE, FUTURE, IRREALIS
TMA FRAME SWITCH	YES, NO, X
VERB CLASS	CLASS 1, CLASS 2, CLASS 3, CLASS 4

Having addressed the predictors associated with the verb and the verb phrase, I turn now to description of the predictor factors related to potential language-external constraints.

5.5.2.4 Predictors related to sociocultural and other language-external factors.

Factors in these categories were either related to individual-specific characteristics, or to aspects of the data collection materials or interviewer(s) that may have had some impact on patterns of SPE.

With respect to data collection materials, although the design and flow of the sociolinguistic interview is intended to simulate casual conversation, the question-and-answer format, and the effects of the observer's paradox (Labov 1972, 1984; Milroy 1980; Wilson 1987), are known to influence speakers' realization of some linguistic forms. Further, the dynamics of turn-taking in conversation and interviews can have a specific impact on SPE, the exact nature of the effect likely differing from language to language as function of pragmatic and cultural norms, though coreference and referent activation likely remain relevant constraints as well. To account for the potential effect on SPE resulting from the question-and-answer format, TURN-TAKING was included as a predictor variable. Outcomes for the predictor TURN-TAKING included RESPONSE, TURN-INTERNAL, and TURN-INITIAL. The latter category indicated when a speaker initiated a new 'discourse chunk' without prompting from the interviewer, usually, though not always, in the picture-description narrative task.

To examine potential differences related to narrative style or data collection materials, the factor TASK was included, with levels INTERVIEW and PICTURE DESCRIPTION NARRATIVE. For most speakers, the picture description task included many more 3rd person referents, and therefore interactions between this factor and PERSON/NUMBER will be considered. Since some interviews were conducted by me alone, and others were conducted with the help of community insiders, the factor INTERVIEWER was included, with levels OUT-GROUP and BOTH. This was to account for any effects that inter-cultural identity relationships may have on SPE for the interviews in which I was the only interviewer present (see Section 5.6, below).

Several factors related to the individual characteristics of the speaker were considered. These were intended to determine if any language-external sociocultural categories might exert some sort of influence on SPE. The factor AGE was a numeric continuous variable. The factors CHILDHOOD DIALECT ZONE and ADULT DIALECT ZONE were included to account for the primary dialect region in which the speaker resided before and during adulthood. Santiago can be split into

roughly three major dialect zones (Quint p.c.; see Section 5.6, below)[72] corresponding to the levels SANTIAGU NORTI, SANTIAGU SENTRU, SANTIAGU SUL; Maio constituted its own dialect zone, DJARMAI (see Image 3 and Section 5.6, below). Cape-Verdeans regularly make sociocultural distinctions, including popular linguistic-stereotypical ones, about residents of Praia, as opposed to residents who live *fora* 'outside', referring to any area outside of the capital. Similarly, the variety of CVC from Santiago and its speakers are typically referred to as *Baidu*, while CVC from other islands and their speakers are collectively known *Sanpadjudu*, or by their island-specific variants. The decision to split dialect zones into childhood and adult stages is due to the fact that many Cabo-Verdeans migrate to Praia for reasons of work, education, general opportunity, and personal relationships. Table 31 displays the distribution of speakers in my sample by their childhood and adulthood dialect zones.

Table 33. Speakers' primary childhood and adulthood dialect regions

Dialect zone	Dialect region in childhood	Dialect region in adulthood
SANTIAGO SUL	12	17
SANTIAGO SENTRU	10	7
SANTIAGO NORTI	4	3
DJARMAI	7	6
Total	33	33

I decided to render this predictor as COMBINED DIALECT REGION, with each dialect zone represented, and one category for speakers who were born and raised in one of the dialect zones other than SANTIAGO SUL, but who reside in Praia in adulthood. The predictor COMBINED DIALECT REGION was thus divided into five groups, four groups were for speakers who resided in the same place they were born and raised, and the other group was for speakers that were from *fora* but live and reside in Praia in adulthood. Thus, the levels for COMBINED DIALECT REGION were: SANTIAGO SUL, SANTIAGO SENTRU, SANTIAGO NORTI, DJARMAI, and COMBINED (SUL + *FORA*).

[72] As has been noted, CVC is rich in dialectal variation. Morphological and phonological variation can be noticeable even from town to town in rural zones.

A final individual-specific sociocultural category was related to socioeconomic status. Based on an adaptation of the metric used in Otheguy & Zentella (2012), a participant's Socioeconomic Status Score (SESSCORE) was calculated by assigning points for education as well as profession (Table 34). For all students who had not yet begun a full-time occupation, their educational scores were doubled.

Table 34. Point values for education and occupation for Socioeconomic Status Score

Education completed	Occupation
1 point – elementary (*Ensino básico*) **2 point** – secondary (*Ensino secundário)* **3 point** – university undergraduate, *'formasôn*' (professional certification) **4 point** – university post-graduate	**1 point** – unskilled (day labor agriculture and construction, odd jobs, panhandling, car washing) **2 points** – blue collar (full time construction, fishermen, full time subsistence agriculture) **3 points** – store owner, primary or secondary public-school teacher **4 points** – white-collar private employee, public functionary

SESSCORE is likely to be highly relevant for CVC, as Cabo Verde is a highly economically stratified society (especially visible in Praia). We have already noted that access to Portuguese, and usually ultimate attainment of Portuguese, are closely related to issues such as access to education, age, and urban vs. rural living. Furthermore, the CVC spoken by an individual can reflect acrolectal or basilectal variants or registers, and all Cabo-Verdeans can modulate their speech to varying degrees in order to emphasize one or the other variant/register. Popularly, acrolectal CVC is referred to as *kriolu levi* 'creole lite' and is loosely associated with upper class, educated, and urban speakers, while *kriolu fundu* 'deep creole' is considered more basilectal and is associated with rural and less educated speakers (see Section 2.6).

The predictive constraints for language-external factors are shown in Table 33, below.

Table 35. Predictors related to language-external factors

Predictor	Levels
TASK	INTERVIEW, PICTURE DESCRIPTION NARRATIVE
TURN	RESPONSE, TURN-INITIAL, TURN-INTERNAL
AGE	numeric continuous, 18-57
COMBINED DIALECT REGION	SANTIAGO SUL, SANTIAGO NORTI, SANTIAGU SENTRU, DJARMAI, COMBINED (*FORA* + SUL)
SESSCORE	Numeric continuous, 2 - 8

5.6 DATA COLLECTION AND SPEAKER PROFILE.

The data for this study were collected over the span of three summers from 2015-2017 on the islands of Santiago and Maio in the Republic of Cabo Verde. The participants were recruited using word-of-mouth snowball sampling, and to the extent possible, the objective was to obtain a diversified cross-sectional sample of CVC speakers representing a range of social categories. In the first summer, I resided in the *Palmarejo* neighborhood of the capital city Praia, the largest urban area of the Republic of Cabo Verde. This neighborhood, popularly identified as middle to upper middle class given its high-rise apartment buildings and modern supermarket facility, is adjacent the *Universidade de Cabo Verde* (UniCV), the sole public university in the country.

During the first summer, most participants were recruited from the UniCV campus, Palmarejo, and surrounding neighborhoods. Several university students were recruited with the help of a UniCV English professor and an English teacher and director of a private English-teaching institute who gave classes to students in a graduate program at UniCV. Other participants

were recruited in the cafes, restaurants, and plazas of Palmarejo and surrounding neighborhoods. Despite the popular socioeconomic conception of Palmarejo, people of all social classes and occupations live in, work in, and transit through the neighborhood. Among those respondents from the first summer who were not UniCV students, they ranged from across the socioeconomic spectrum. Further, many UniCV students move to Praia from rural areas of Santiago to attend university, living with relatives or friends. During a five-day excursion to the island of Maio, respondents were recruited in the municipality of *Vila do Maio*, also known as *Porto Inglês.*

In total, I interviewed 31 speakers in the first summer, 20 of whom were residents of Praia. 11 interviews were retained from Praia-based speakers for the final analysis, of these speakers, three were originally from *Santiago centro*, one was originally from *Santiago norte,* and one was originally from *Maio*. Of the 7 interviews collected from residents of Maio, five were retained for analysis; one speaker that was based in Maio for most of his adulthood grew up in *Santiago centro* and travels there once a month.

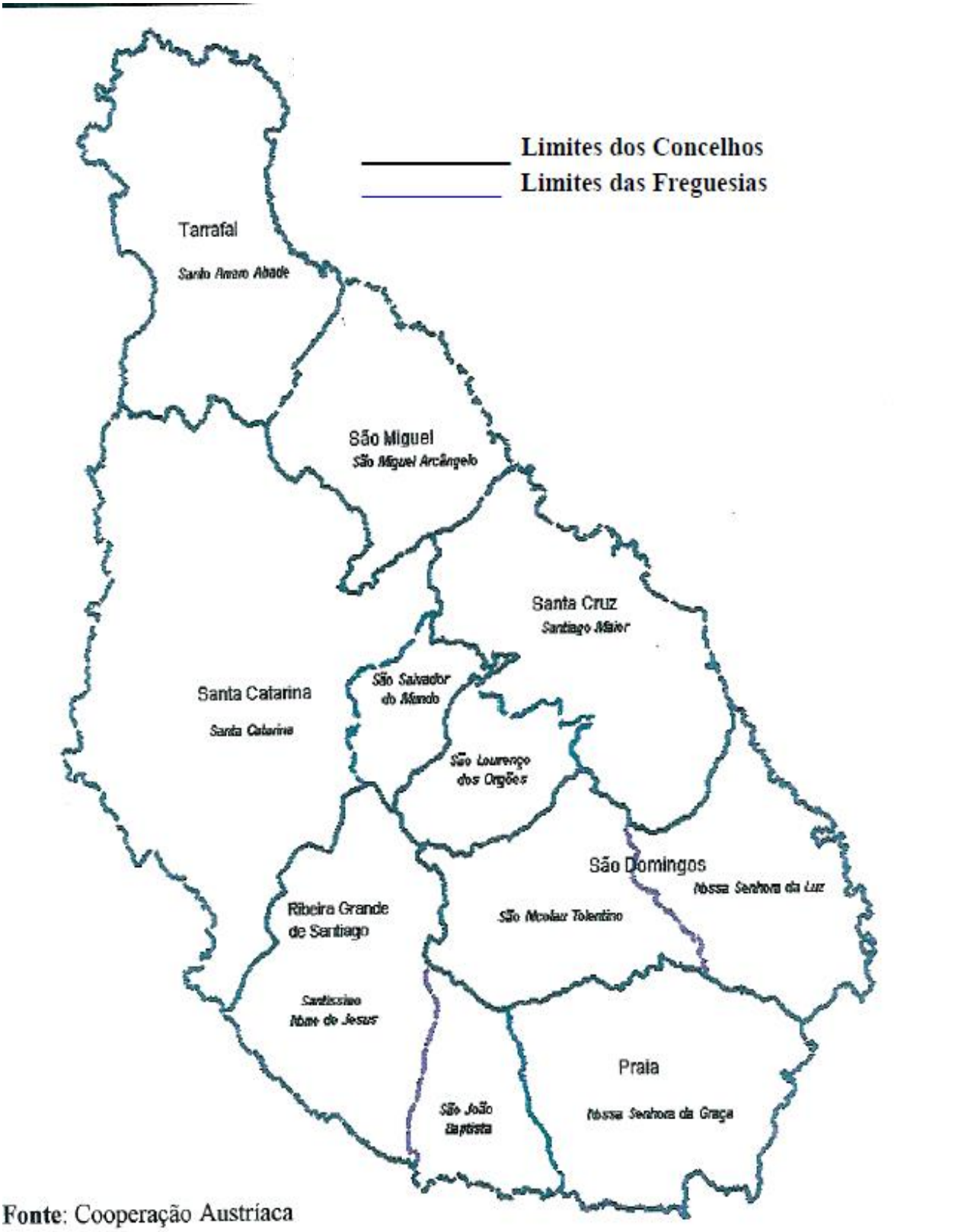
During the second summer, I resided in the *Achada Igreja* urbanization of the *vila* 'town' *Picos*, which is the municipal seat of the rural and dispersedly populated *concelho* 'municipality' *São Salvador do Mundo,* located in the mountainous center of Santiago. Participants were recruited from throughout *São Salvador do Mundo,* as well as the adjacent *São Lourenço de Órgãos*, and from semi-rural *Santa Catarina,* which has its municipal seat in *Assomada,* the second largest city on Santiago and fourth largest in Cabo Verde. In the second summer, most interviews were carried out with the assistance of a friend, colleague, and confidant, who is a native speaker of CVC, works and attended university in Assomada, and was born and raised in Achada Igreja. This colleague was also essential in helping me to meet and make friends in the local community, and in identifying and recruiting participants for the interview and picture description tasks. In total, in the second summer, we collected recordings from 22 speakers, ten of these were retained for analysis in the present study.

In the third and final summer, I resided in the beachside town of Tarrafal in the eponymous *concelho* in northern Santiago. Just as in the first summer, a local colleague and confidant who is

a native speaker of CVC and who was born and raised in Tarrafal aided and assisted in both identifying and recruiting participants, in asking questions during the interview task, and explaining instructions for the picture-description narrative task. In total, in the third summer, we collected recordings from 27 speakers, eight of these interviews were retained for analysis.

The delimitation of the regions on Santiago into *sul, centro*, and *norte* reflected broad dialectal zones (Image 3) (Quint p.c.). *Santiago norte* corresponded with the *concelhos* of *Tarrafal* and *São Miguel*, *Santiago centro* with *Santa Catarina, São Salvador do Mundo, Santa Cruz,* and *São Lourenço dos Órgãos*, while *Praia, São Domingos,* and *Ribeira Grande de Santiago* were considered *Santiago sul.* In the interviews, participants were asked about where they grew up and how long they had been living in their current place of residence. Based on their responses, participants were classified by their primary childhood dialect region and their primary adulthood dialect region (Table 32, Section 5.5.2.4).

Image 3. Map of the *concelhos* 'municipalities' and *freguesias* 'parishes' of Santiago, Cabo Verde (Ramos Moreira 2005:"source: *Cooperação Austríaca*")



Fonte: Cooperação Austríaca

Interviews ranged from 12 to over 90 minutes in length and were recording using a Zoom H2N Handy recorder. All interviews consisted of a semi-spontaneous guided sociolinguistic interviews. These interviews were designed to be conversational in nature but were structured by a set of questions selected to elicit a wide range of topics, themes, types of subject referent, and Tense/Mood/Aspect frames. Each interview relied on roughly the same set of questions, but if the interviewer(s) found a topic to be of interest to the respondent, then the they pursued additional questions on that topic or allowed the respondent to continue as they saw fit. 25 of 33 respondents also completed a picture-description task in which they were asked to narrate the progression of events depicted in *The Frog Story* (Mayer 1967), a task that has been used in a range of linguistic and psychological studies (cf. Sánchez 2003; Slobin 2004; *inter alia*). This task also served to elicit more 3rd person referents. Speakers ranged from age 18 to 57, and the average age was 29.3 years old. There were 12 male informants and 21 female informants. The average age of the female informants was 25.4 and the average age of the male informants was 36.2. There were two participants with an SESSCORE of 2, one with a score of 3, five with a score of 4, three with a score of 5, fifteen with a score of 6, and seven with a score of 8. The profile of the study participants is summarized in Table 36.

Table 36. Participant profile

Paritic pant code	Sex	Combined Dialect Region	Age	Education completed	Employment/current education	SES SCORE
P1	F	SUL	24	Undergraduate	Graduate school	8
P2	M	SUL	33	Undergraduate	Graduate school	8
P3	M	NORTI	30	Secudário	Full time construction	4
P4	F	SUL	27	Undergraduate	Graduate school	8
P6	F	SUL	28	Undergraduate	Graduate school	8
P8	M	*FORA* + SUL	21	Básico	Day labor, odd jobs	2
P9	F	SUL	19	Secudário	Undergraduate	6
P10	M	SUL	37	Secudário	Grocery store clerk	3
P11	F	SUL	27	Secudário	Undergraduate	6
P12	F	SUL	18	Secudário	Undergraduate	6
P13	F	*FORA* + SUL	18	Secudário	Undergraduate	6
P14	F	SUL	23	Secudário	Undergraduate	6
P15	F	SUL	19	Secudário	Undergraduate	6
P16	M	*FORA* + SUL	19	Secudário	Undergraduate	6
P17	F	SUL	18	Secudário	Undergraduate	6
P18	M	SUL	43	Secudário, *formasôn*	Undergraduate, secondary teacher	6
P19	F	*FORA* + SUL	21	Secudário	Undergraduate	6
P20	F	SUL	22	Secudário	Undergraduate	6
P21	F	DJARMAI	21	Secudário	Convenience store clerk	5
P23	M	DJARMAI	28	Secudário	Full time construction	4
P24	M	DJARMAI	55	Secudário	Store owner	4
P25	F	DJARMAI	24	Secudário	Convenience store clerk	5
P26	F	DJARMAI	24	Secudário	Convenience store clerk	5
P27	F	DJARMAI	24	Undergraduate	Bank teller	6
P28	M	SENTRU	32	Básico	Day labor, odd jobs	2
P32	M	SENTRU	55	Secudário, *formasôn*	Retired primary teacher	6
P34	F	NORTI	28	Secudário	Undergraduate	6
P38	M	SENTRU	51	Secudário, *formasôn*	Retired	4
P41	F	SENTRU	46	Básico	Full time subsistence agriculture	3
P45	F	SENTRU	21	Secudário	Full time subsistence agriculture	4
P46	F	SENTRU	57	Secudário	Full time subsistence agriculture	4
P47	F	SENTRU	25	Secudário	Undergraduate	6
P53	M	NORTI	30	Secudário, *formasôn*	Day labor, odd jobs	4

5.7 STATISTICAL PROCEDURE.

In this final subsection of the Chapter 5, I briefly outline the statistical procedures applied in the present study. These procedures are described in more detail in Chapter 6 prior to conducting a technical description of the results of the analysis. Descriptive and inferential statistical analysis was conducted in R (R Core Team 2019). The data were examined using conditional inference trees (Hothorn, Hornik, & Zeilis 2006), particularly to explore interactions among predictor variables. Random forests were grown and visualized in variable importance plots to assess the permutation feature importance measure of the predictor variables (Hothorn *et al.* 2006; Strobl *et al*. 2007; Strobl *et al.* 2008). Conditional inference trees are a type of decision tree, specifically a non-parametric regression tree, and random forests rely on an ensemble of conditional inference trees. These procedures also guided the selection of constraints for inclusion in the regression models.

I constructed four different analyses to examine the data descriptively and inferentially. The first analysis, Analysis 1, was fully exploratory; it was used to construct subsequent analyses, but aside from that, only the descriptive observations from Analysis 1 (but not the inferential observations) directly informed the ultimate discussion of the results. The purpose of Model 1 was to conduct a preliminary examination of correlations between dependent and predictor variables for the purpose further delimiting the envelope of variation that was applied to subsequent analyses. This first analysis also involved a three-way response variable: SC, X2SBJ, Ø, requiring a multinomial logistic regression for inferential examination, and Analysis 1 also represented the broadest conception of the envelope of variation. Subsequent analyses removed and merged contexts of non-variability associated with certain predictor or levels of the predictor variables.

For the inferential analysis, the data were submitted to multinomial logistic regression (MLR, henceforth) using `nnet` (Venables & Ripley 2002). MLR is applied when dealing with response variables with more than two discrete outcomes. When compared to running two or more binomial logistic regression models with the levels of the multinomial response variable merged or split into pairs, MLR reduces the probability of committing Type 1 Error (false positive), or

Type 2 Error (false negative), and increases power (Pituch & Stevens 2015:1-41). This procedure formed the basis for the preliminary exploratory analysis, Analysis 1. After descriptive and inferential observation of Analysis 1 and further delimitation of the envelope, Analysis 2 was also constructed using MLR (see Rosemeyer & Enrique-Arias 2016 and Rosemeyer & Schwenter 2019 for application of MLR to linguistic corpus data). Analysis 2 was also inspected for multicollinearity and singularity.

Analyses 3 and 4 involved binomial logistic regression models that merged one of the response variable outcomes with the SC-level for comparison. Analysis 3 compared Ø subjects against both overt subject forms, while Analysis 4 compared X2SBJs against Ø subjects and SCs.

The model building and factor selection process for the binomial models, in addition to relying on random forests to assess variable importance, also relied on the `step` function to conduct stepwise ANOVA for the comparison of nested regression models based on the Akaike Information Criterion (AIC) (cf. Baayen 2008; Tagliamonte & Baayen 2012). Predictors associated with the most variable importance from the random forest procedure, and that were retained in nested models with low AIC scores, were used as base-models against which to compare similarly constructed models with comparable combinations of parameters. The base-model, and the similar models reflecting slight modifications to the combination of parameters in the base-models, were again compared for their AIC scores before selecting a final model (this latter procedure was also applied to the MLR in Analyses 1 and 2). Both binomial models were also inspected for multicollinearity and singularity. Descriptive results were plotted using `ggplot2` (Wickham 2016) and the output of the inferential models were plotted using `sjPlot` (Lüdecke 2019). The statistical procedures are described in more detail for each model in the next chapter.

Chapter 6: Results

This chapter presents the results from descriptive and inferential analyses of patterns of Subject Pronoun Expression (SPE) in the corpus of Cabo-Verdean Creole (CVC) described in the previous chapter. The objective in this chapter is to provide a technical description of the results, which is followed by a broader discussion and framing in the following chapter in which I draw conclusions on their broader relevance for themes of the dissertation. It was explained that since this is the first study to attempt to delimit the envelope of variation for SPE in CVC, I began with the broadest possible definition of the envelope for [+referential, +argumental] target anaphoric subjects (see Section 5.4). The relevant nominative anaphora include a subject clitic (SC), a zero subject/anaphoric zero (Ø), and a double-subject construction (X2SBJ) involving a tonic pronoun and a SC. The predictors considered were elaborated in Section 5.5, and hypotheses for how various constraints are assumed influence SPE were considered in 5.4.

Crucially, SCs and Ø compete to resume antecedents with animate, definite, and specific reference, as well as antecedents with the opposite properties, and it is a matter pending empirical confirmation whether X2SBJs can too (Section 5.4). Therefore, lexical DPs with inanimate, indefinite, or nonspecific reference could also serve as possible antecedents for anaphora eligible for inclusion within the envelope of variation (this also applied to the DP + INTERVENING MATERIAL + Ø construction; see Section 5.4 and 5.5.2.1). With this initial broad-envelope, a full descriptive and inferential analysis is conducted involving the procedures for coding predictors described in Section 5.5.2. This primary analysis will be used to further constrain the envelope of variation and build the design for subsequent analyses (Section 5.7).

The first step is aimed at identifying contexts of invariability or low variability with respect to the use of all three anaphoric subject forms and their correlating predictors. The motivation for this is multifold; to begin with, it is a methodological requirement of a variationist analysis that all observations of the response variable outcomes be recorded across congruent morphosyntactic and discursive contexts. Since the morphosyntactic and discursive distribution of subject anaphora are often contingent on their relationship to their antecedents and the properties of antecedents, as well

as their local environment, the actual distribution of subject anaphora relative to the predictors that encode these properties must directly inform an empirical delimitation of the variable context. Low token counts for a response outcome for levels of a predictor variable can also be a source of collinearity, increased error, and other issues when attempting to fit a regression model.

Further, it is well known that non-argumental or quasi-argumental constructions like impersonals, existential, and presentationals, pattern differently from genuinely referential nominative arguments with respect SPE and anaphora resolution cross-linguistically (Svenonius 2001; Kaiser 2015). By restricting the range of anaphoric expressions considered to those that have a discernable antecedent, even antecedents that are to some degree referentially deficient in their semantic value, we can tease apart the distributional patterns of referential anaphora from non-argumental or quasi-argumental subject constructions (Section 5.4).

There were two predictors that were coded for but nonetheless excluded from model selection, TMA FRAME and SEX. The former described the TMA reading associated with a given verb phrase for a target subject, since there is no one-to-one correspondence between a TMA marker and its actual temporal/modal/aspectual reading in CVC, and because bare verbs can also be tensed but differ in their temporal and aspectual interpretation depending on their stativity (see Sections 5.4 and 5.5.2.3). These predictors were excluded because there were no specific predictions associated with them as far as SPE in CVC is concerned, neither from the previous literature on CVC, nor from the literature on other languages that were surveyed in preceding chapters. Another reason for the exclusion of these predictors was the already large number of parameters under consideration, which were too many to all be included in a convergent model without issues of collinearity and without detriment to model fitness.

The first subsection of this chapter describes Analysis 1. Recall that Analysis 1 is exploratory and used to inform the construction of subsequent analyses (Section 5.7); the descriptive results from Analysis 1 inform the ultimate discussion and conclusions drawn, but the inferential statistical output from Analysis 1 does not. Instead, it serves to gain a preliminary understanding of the distribution of the response variable as it correlates with predictors before

constraining the model further. Visualizations of the descriptive and inferential results are nonetheless presented side-by-side for each predictor. I believe this presentation allows for a clear picture of the nature of the variable before it is delimited further.

The next subsection describes the second analysis, Analysis 2. This analysis also considers the three-way response variable (SC, Ø, X2SBJ), but incorporates delimitations to the envelope of variation, including the removal of Ø anaphora with nonpast verbs from CLASSes 3 and 4 (Section 5.4), and the merger or exclusion of contexts found to be invariable or nearly-invariable in Analysis 1. As with the first analysis, the descriptive results are presented alongside the results of the inferential analysis where relevant. A variable importance plot of a random forests model is presented, along with conditional inference trees, the latter were particularly useful for inspecting interactions among the predictors. The inferential analysis is conducted using a fixed-effects multinomial logistic regression (MLR). I also assess the MLR in Analysis 2 for multicollinearity and singularity, concluding that is a well-fitted model with low levels of multicollinearity.

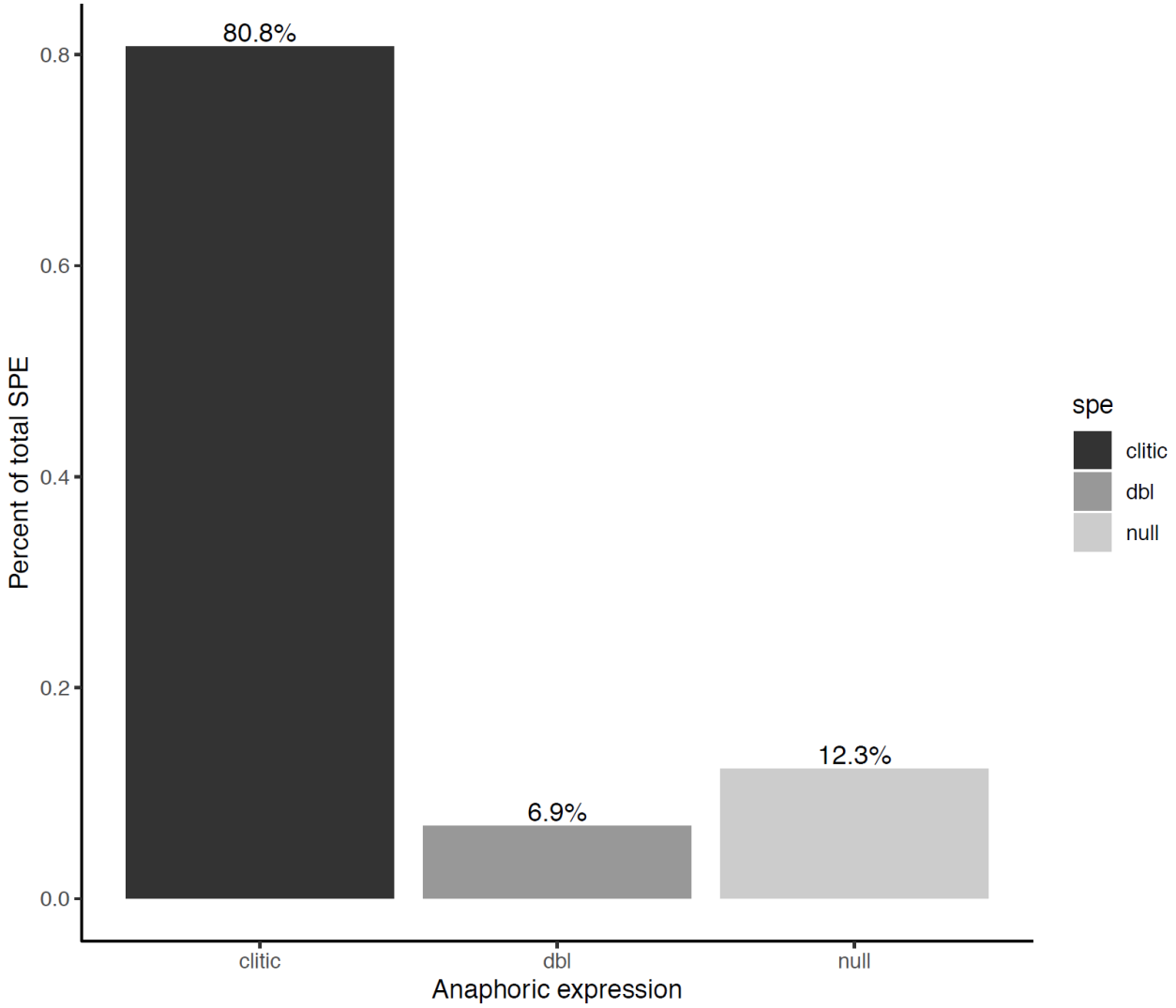
The final subsections of this chapter described Analyses 3 and 4. These analyses consisted of mixed-effects binomial logistic regressions. These models were built in order to account for the variance related to random effects such a speaker (PARTICIPANT), and to confirm the results from Analysis 2. Each model involved a merger of one of the levels of the three-way response variable. The regression model in Analysis 3 compared Ø subjects against the OVERT realizations (SCs and X2SBJs), while the regression model in Analysis 4 compared X2SBJs against a merged category for all other subject anaphoric realizations, ALL ELSE (SCs and Ø). For each analysis, I also described the model building and parameter selection procedure. Predicted probabilities and conditional inference trees were used to examine the combined effects of predictors.

6.1 ANALYSIS 1: MULTINOMIAL LOGISTIC REGRESSION MODEL 1, BROADEST ENVELOPE OF VARIATION.

The first analysis, Analysis 1, is built upon a variable context following the design described above and throughout the previous chapter. Recall that the restrictions on admitting Ø subjects targets with BARE verbs from VERB CLASSes 3 and 4 are not applied until Analysis 2, meaning that all target anaphora with BARE verbs from VERB CLASSes 3 and 4 were admitted for analysis in Analysis 1 (see Section 5.7).

Approximately 8,500 observations were isolated; most of these contained at least one Intonational Unit (IUs) (Sections 4.5.3 and 5.5.1). For IUs with several verbs, each verb was isolated in a distinct, individual observation along with its corresponding subject 'slot' (line in the Excel spreadsheet). After eliminating contexts excluded from the envelope following the procedure described in Sections 5.4 and 5.5, a total of 3,995 anaphoric subject observations were retained. The proportions of Ø, SCs, and X2SBJs are shown in Figure 9.

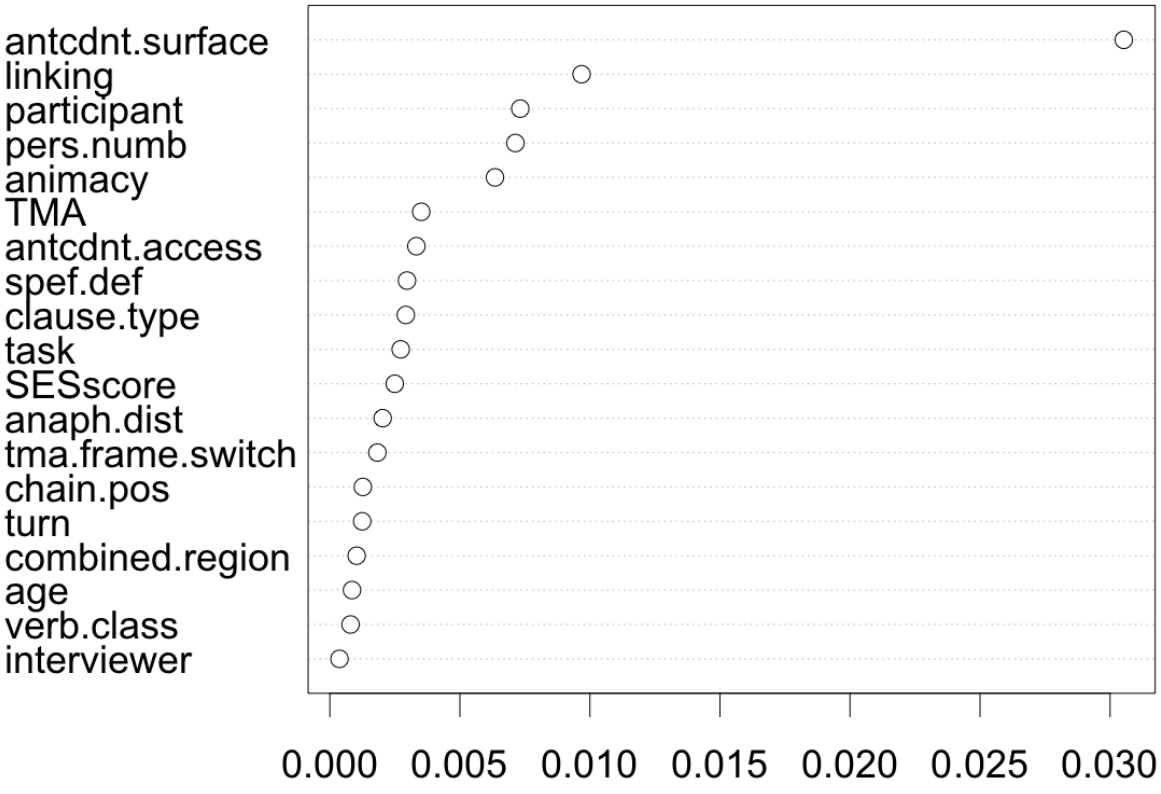
Figure 9. Overall rate of each anaphoric subject type in a corpus of CVC, Analysis 1.



As predicted, SCs were by far the most productive subject anaphor at 80.8%, while Ø subjects were the second most frequent at 12.3%. This is slightly higher than what Duarte & Soares da Silva (2016:16) found for second person subjects in Brazilian Portuguese (BP), though much lower than the BP rate of anaphoric Ø for 3rd person referents (=41%). X2SBJs accounted for 6.9% of all subject realizations, close to the rates of overt subjects in the 3rd person that Duarte & Soares da Silva found for Italian (2016:7).

A variable importance plot of a random forest procedure for Model 1 is shown in Figure 10. This plot includes the factor PARTICIPANT, which will be included as a random factor in later models. This was done in order to assess the degree of variation that could be explained by the idiosyncratic patterns of SPE produced by the study participants alone. In the next figure, Figure 11, PARTICIPANT is not included as a parameter in the random forest procedure. The two variable importance plots were then compared for differences. In these plots, each predictor variable is listed in descending order according to its relative importance in the model in terms of the extent of the variation associated with SPE that can be explained by each factor. Dots plotted up and furthest right are the most important and explain the greatest proportion of the variation in the model.

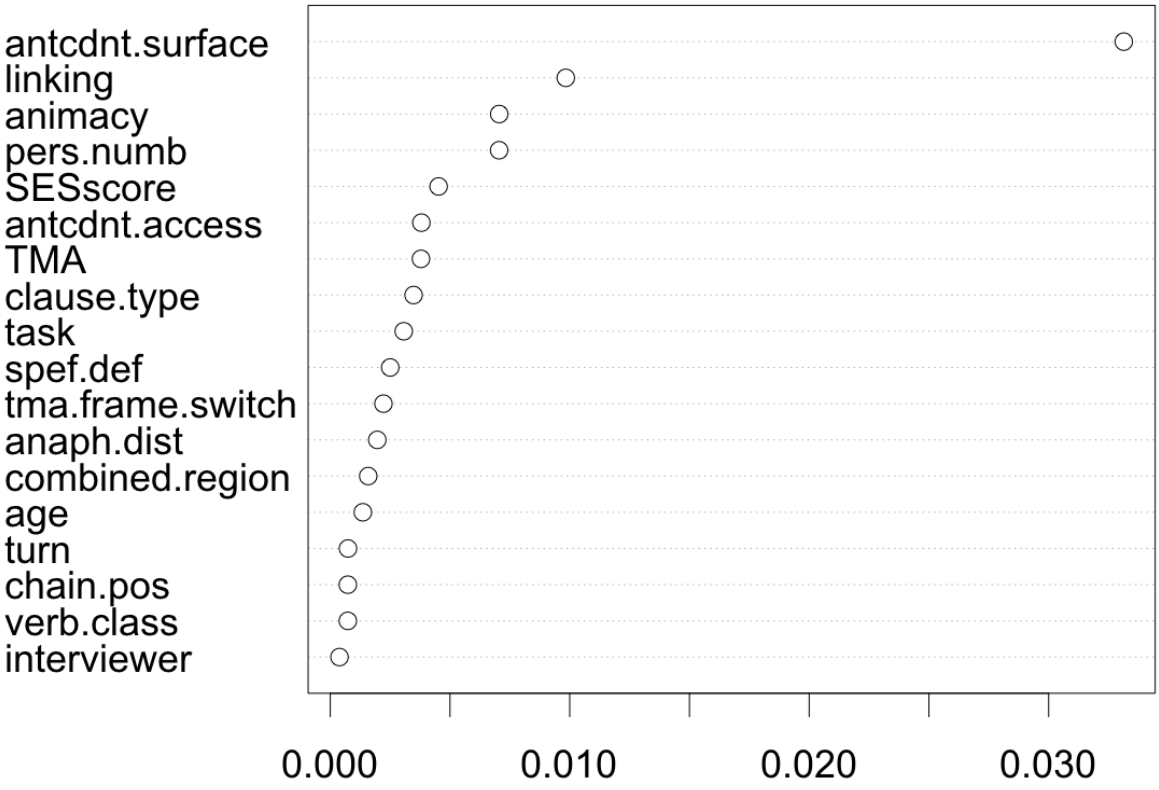
Figure 10. Variable importance plot of random forest for all predictors in Analysis 1, broadest envelope of variation (including PARTICIPANT).



In Figure 10 one can see that SURFACE FORM OF THE ANTECEDENT (Section 5.5.2.1) was by far the most important predictor in the model and can be assumed to explain the largest proportion of variability due to its high Gini importance (x-axis values). This was followed by a large mean decrease in Gini importance associated with the next predictor factor, LINKING, followed by another drop-off, and a group of three factors of similar Gini importance, PARTICIPANT, PERSON/NUMBER, and ANIMACY (Section 5.5.2.2). The large amount of variation that be explained by the category PARTICIPANT indicates that these data would benefit from the application of mixed-modeling that accounts for the variation attributable to individual-specific patterns of SPE. After ANIMACY, there is another drop-off and the next most important predictor factor is TMA, followed by ANTECEDENT ACCESSIBILITY PATTERN, and another slight drop-off to SPECIFICITY/DEFINITENESS, CLAUSE TYPE, TASK, SESSCORE, ANAPHORIC DISTANCE, TMA FRAME SWITCH, and COMBINED DIALECT REGION, and so forth.

In Figure 11 the random forest is grown once again, this time excluding PARTICIPANT. The variable importance ranking in the previous plot (Figure 10 above) is then compared with the ranking that excluded PARTICIPANT.

Figure 11. Variable importance plot of random forest for all predictors in Analysis 1, broadest envelope of variation (excluding PARTICIPANT).



The results for the variable importance plot for the random forests procedure excluding PARTICIPANT in Figure 11 shows no differences in the order or ranking of predictor variables before (above) PARTICIPANT listed in Figure 10 (the first two predictors are SURFACE FORM OF THE ANTECEDENT and LINKING). The removal of PARTICIPANT caused the predictor ANIMACY to rise in importance, now coming third after LINKING, followed by PERSON/NUMBER. SESSCORE now surpassed ANTECEDENT ACCESSIBILITY PATTERN, which itself rose in importance, now coming before TMA. After this came CLAUSE TYPE, TASK, and SPECIFICITY/DEFINITENESS, followed by TMA FRAME SWITCH, ANAPHORIC DISTANCE, COMBINED DIALECT REGION, and so forth.

Predictor factors were entered into the model in the order reflecting their ranking in the second variable importance plot. The first model included all fixed effects predictors, but when this model did not converge (due to too large of a number of parameters in proportion to token count relative to overall occurrence of the non-baseline forms (Ø and X2SBJs), or the inclusion of non-explanatory variables, or some combination of these issues), variables were removed one-by-one until convergence (backwards selection). After a convergent model was obtained, it was used as a 'base' model upon which to construct similar models for comparison.

The subsequent iterations of the convergent 'base' model contained different combinations of factors and interaction terms. Using the `step` function in R (R Core Team 2019), nested regression models were compared by stepwise algorithm. Under this procedure, models are attributed an Akaike Information Criterion (AIC) score, which relies on the chi-square goodness of fit test to penalize the addition of a parameter and weigh the penalty against the ability of that parameter to explain variance in the model; the model with the best AIC score is that which obtains the best fit with smallest number of parameters (the most variance explained with the lowest degrees of freedom). This procedure was repeated for each permutation of the envelope considered in Analyses 1-4. After selecting the best model using this process, the log odds of the model estimates and their standard deviations were plotted using `sjPlot` (Lüdecke 2019). The numeric model output tables from which the plots are constructed are presented in appendices (2, 3, 5, and 7). Descriptive plots were built using `ggplot2` (Wickham 2016); these are presented alongside the corresponding inferential outputs for several of the statistically significant predictors. I also provide some plots of the predicted probabilities for realizing each anaphoric subject form as an effect of combinations of two or three independent variables to examine their combined effects (`sjPlot`, Lüdecke 2019).

For the converged model of best fit associated with the broadest envelope (Analysis 1), the factors ultimately selected were entered in the following order: SURFACE FORM OF THE ANTECEDENT, LINKING, ANIMACY, PERSON/NUMBER, TMA, SESSCORE, ANTECEDENT ACCESS

PATTERN, CLAUSE TYPE, TASK, SPECIFICITY/DEFINITENESS, TMA FRAME SWITCH, ANAPHORIC DISTANCE, and COMBINED DIALECT REGION[73] (Appendix 3).

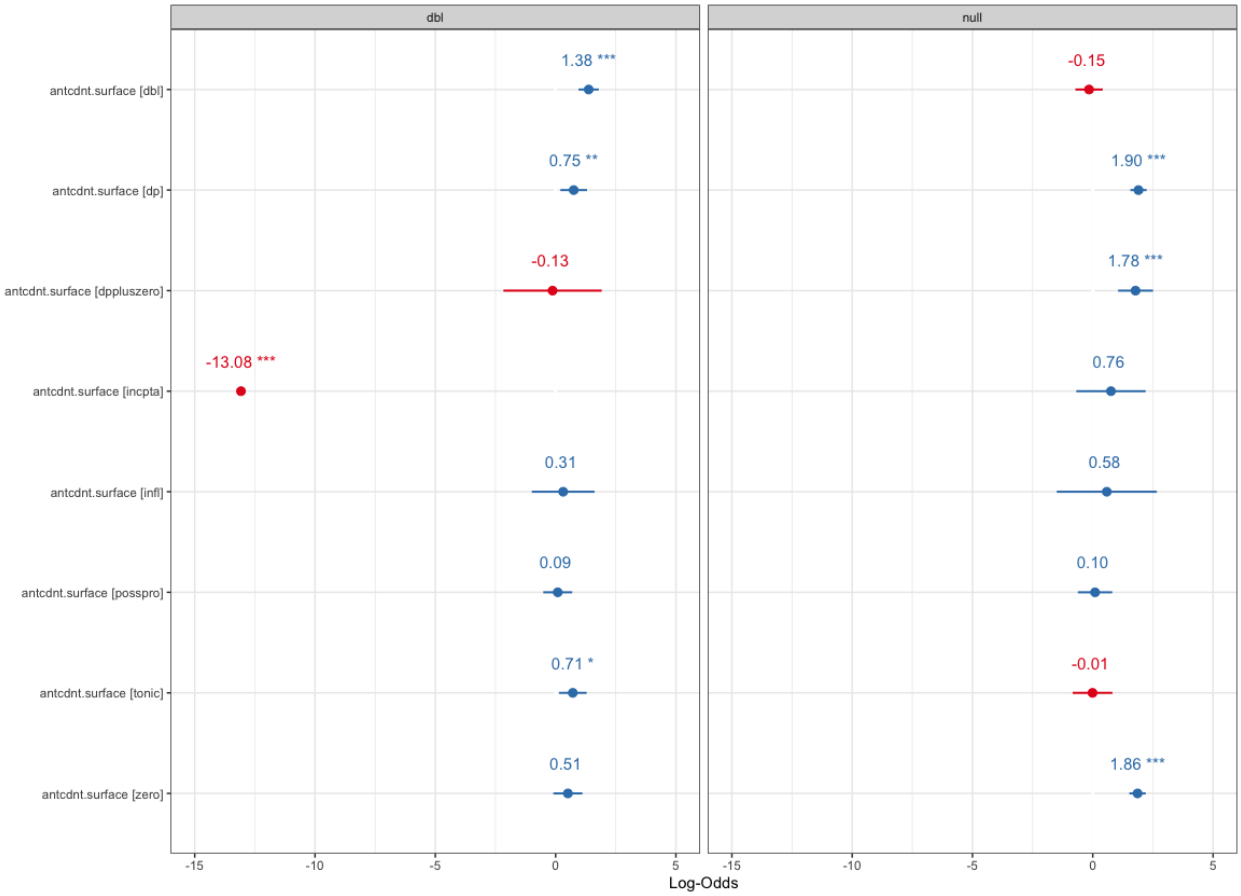
SURFACE FORM OF THE ANTECEDENT described the antecedent's morphological category (part of speech) (Section 5.5.2.1). It was included to examine the effects of linguistic structural priming on SPE, as well as other discourse organizational effects related to antecedent accessibility. This predictor was returned as the most important factor in the variable importance plot. Correspondingly, it was highly significant in the multinomial logistic regression model, where it was shown to exert strong effects on SPE outcomes.

The plot in Figure 12 is the first in a series of visualizations of the multinomial logistic regression output. It displays the log-odds (LODs) of realizing a Ø subject (right-hand panel) or a X2SBJ (left-hand panel) over a SC, respectively, when the SURFACE FORM OF THE ANTECEDENT was one of the outcomes on the y-axis (predictor levels: X2SBJ, LEXICAL DP, DP + INTERVENING MATERIAL + Ø, INCORPORATED *TA* MEDIAL SUBJECT, INFLECTION, POSSESSIVE PRONOUN, TONIC, Ø), as opposed to the reference-level value, which for this predictor was CLITIC.

The LODs are indicative of the magnitude of the effect, with values further away from zero indicating a stronger effect. LODs to the left of zero have negative values and are associated with a disfavoring effect on the response variable outcome in its corresponding panel, while those to right of zero have positive values and are associated with a favoring effect. The asterisks represent a significant p-value (* = $p < 0.05$, * = $p < 0.01$, *** = $p < 0.001$.). The error bars intersecting each plotted point represent the standard error associated with that parameter, an estimate of the standard deviation of the regression coefficients.

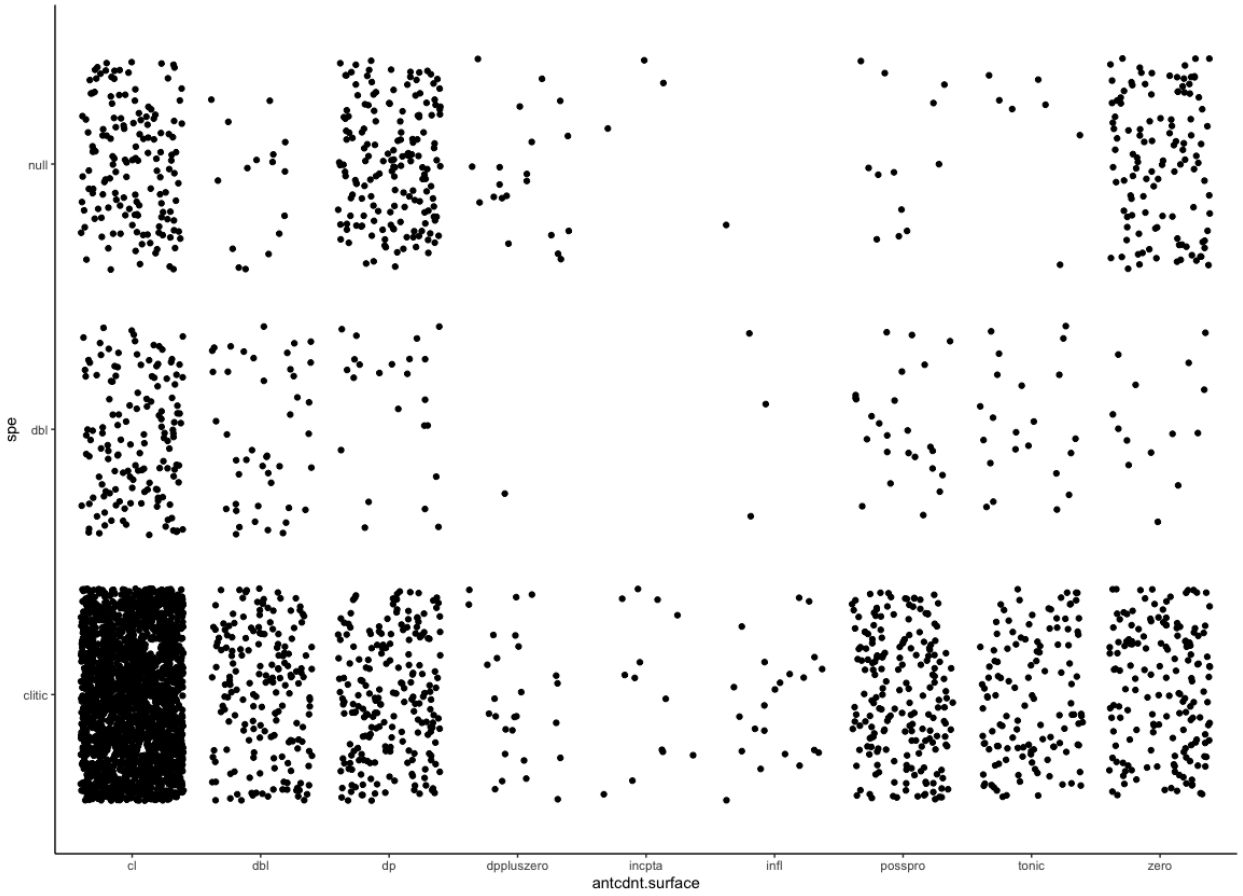
[73] AIC = 3835.01, Residual Deviance = 3627.005

Figure 12. The effects of SURFACE FORM OF THE ANTECEDENT (application value = CLITIC) on SPE, MLR1, Analysis 1, broadest envelope.



A descriptive plot of the distribution of subject forms by SURFACE FORM OF THE ANTECEDENT is shown in Figure 13. Each plotted point represents a subject realization, with each outcome plotted on the y-axis (Ø, X2SBJ, SC) and each level of the predictor plotted on the x-axis (predictor levels: CLITIC, X2SBJ, LEXICAL DP, DP + INTERVENING MATERIAL + Ø, INCORPORATED *TA* MEDIAL SUBJECT, INFLECTION, POSSESSIVE PRONOUN, TONIC PRONOUN, Ø).

Figure 13. Jitter plot of SPE x SURFACE FORM OF THE ANTECEDENT, Analysis 1, broadest envelope.



One can observe that several different levels of the constraint SURFACE FORM OF THE ANTECEDENT were highly predictive of SPE outcomes. The first significant result indicates that a X2SBJ antecedent, as opposed to a CLITIC antecedent, exerted a favoring effect on the realization of a X2SBJ as the target subject (LOD = 1.38, $p < 0.001$), when compared to a SC as the target subject. In other words, there was a structural priming effect for X2SBJs; the realization of a X2SBJ in an antecedent had a favoring effect on the realization of a X2SBJ in a target.

The same effect can be found for TONIC pronouns antecedents (monosyllabic and disyllabic; oblique, subject, or object); when compared to CLITIC antecedents, TONIC pronoun antecedents favored a X2SBJ realization (LOD = 0.75, $p < 0.05$), as opposed to a SC. This suggests

that the prosodic, discourse, and syntactic status of lone TONIC pronouns is comparable to that of X2SBJ constructions, at least in so far as structural priming effects are concerned.

Ø antecedents also exerted a structural priming effect; a Ø antecedent, as opposed to a CLITIC, was predictive of a Ø target subject (LOD = 1.86, $p < 0.001$), when compared to a SC. The putative (medial) subject position of an INCORPORATED *TA* construction (see Section 5.4), as opposed to a CLITIC (in some other configuration), exerted a strong disfavoring effect on the realization of a X2SBJ (LOD= -13.08, $p<0.001$), when compared to a SC. Indeed, as can be seen in the jitter plot in Figure 13 above, X2SBJs never occurred when the antecedent was the putative (medial) subject position of an INCORPORATED *TA* construction.

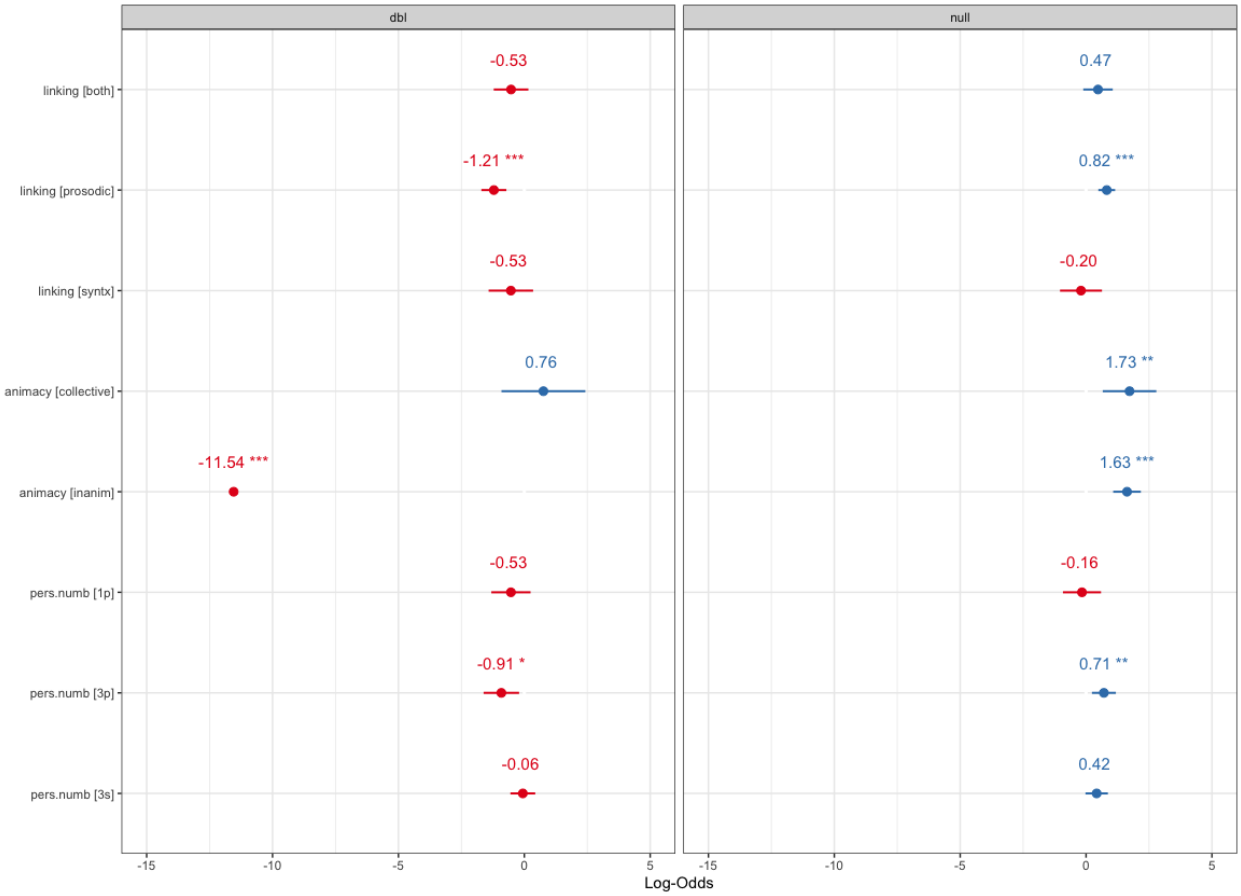
Another set of findings related to the SURFACE FORM OF THE ANTECEDENT subject are not attributable to priming effects. A LEXICAL DP antecedent, as opposed to a CLITIC, was predicative of realizing both X2SBJs (LOD = 0.75, $p < 0.01$), and Ø subjects (LOD = 1.90, $p < 0.001$), when compared with a SC target subject, respectively. Part of the effect from LEXICAL DP antecedents is likely due to their high degree of topicality and prominent prosodic/discourse status. Such antecedents are highly activated and silent in the discourse, which enhances the degree to which they are 'referentially continuous' and accessible to subsequent anaphora (see Section 4.5.3). Under circumstance of high referential continuity and accessibility, a Ø subject is more likely to occur, presumably because the processing burden associated with referential recovery is minimal (cf. Givón 1983a; 2017; Ariel 1990). As will be shown ahead, however, this relationship between a LEXICAL DP antecedent and Ø is active at short anaphoric distances. It is also associated, to certain extent, with a set of semantic properties of the antecedent subject referent. On the other hand, as we will see, the relationship between LEXICAL DPs and X2SBJ targets is active at longer anaphoric distances, where X2SBJs serve to re-introduced old discourse referents and establish contrast when switching between two or more referents.

The same relationship that was found for LEXICAL DPs and Ø can be extended to the DP + INTERVENING MATERIAL + Ø construction (see Section 5.4). When the DP + INTERVENING MATERIAL + Ø construction was an antecedent, as opposed to a CLITIC, this context was also favoring

condition for a Ø target subject (LOD = 1.78, $p < 0.001$), over a SC. We will see that the selective effect of the DP + INTERVENING MATERIAL + Ø antecedents on Ø targets is also related to the activation of discourse referents, as well as the referents' semantic properties. This indicates that the DP + INTERVENING MATERIAL + Ø construction is equivalent to a (verb phrase adjacent) LEXICAL DP, at least in so far as antecedent-anaphor relationships are concerned. After presenting the results for the predictors ANIMACY and SPECIFICITY/DEFINITENESS, I will return to discuss the findings for LEXICAL DP and DP + INTERVENING MATERIAL + Ø antecedents.

The next most important constraints in Analysis 1 were LINKING (predictor levels: PROSODIC, SYNTACTIC, BOTH, NO LINK) (Section 5.5.2.1), ANIMACY (predictor levels: ANIMATE, COLLECTIVE, INANIMATE) (Section 5.5.2.2), and PERSON/NUMBER (predictor levels: 1SG, 1PL, 3SG, 3PL) (Section 5.5.2.2) (Figure 14).

Figure 14. The effects of LINKING (application value = NO LINK), ANIMACY (application value = ANIMATE), and PERSON/NUMBER (application value = 1SG) on SPE, MLR 1, Analysis 1, broadest envelope.



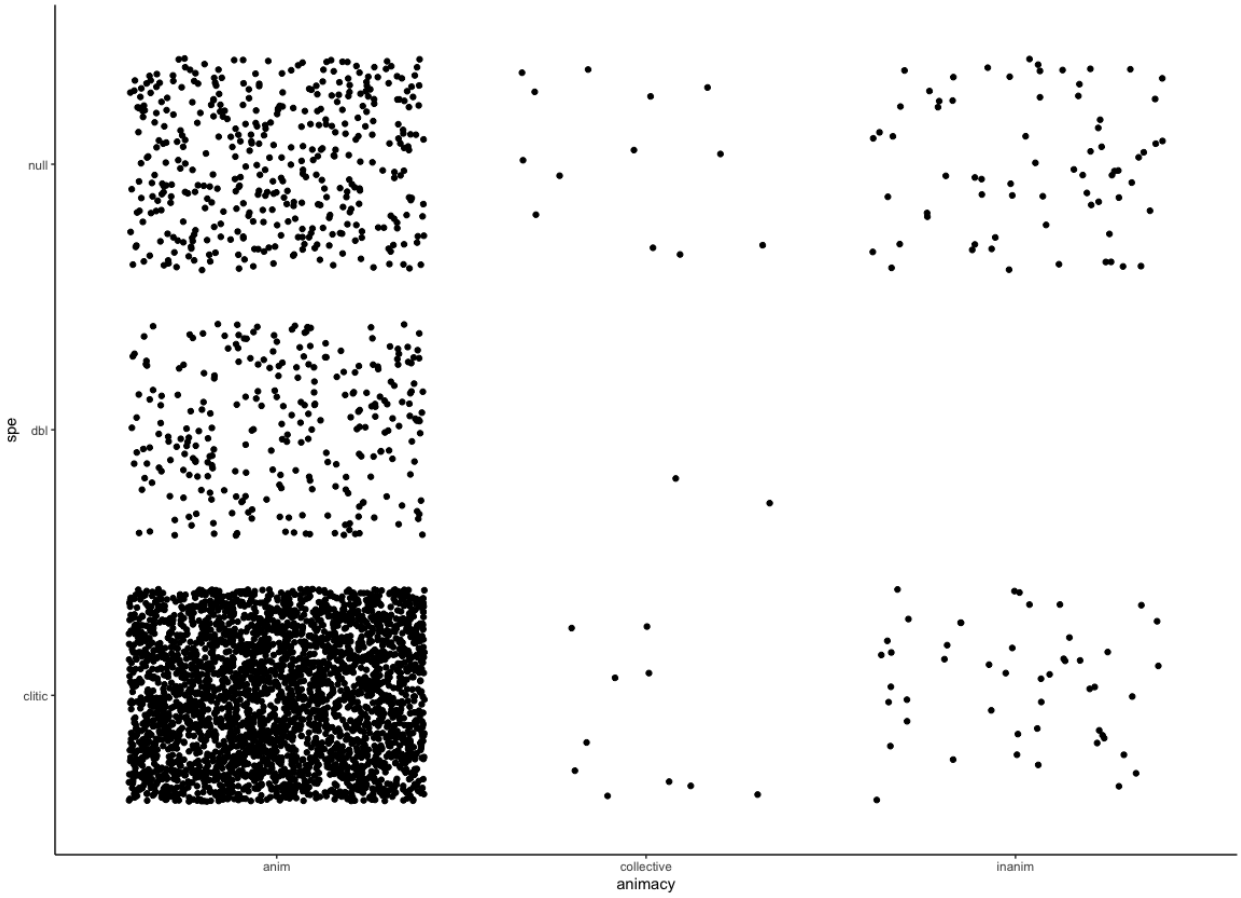
PROSODIC LINKING, as opposed to NO LINK, disfavored the realization of a X2SBJ (LOD = -1.21; $p < 0.001$), but favored the realization of Ø (LOD = 0.82; $p < 0.001$), when these were compared to a SC, respectively. This is consistent with what has been found for other languages: Ø is promoted by prosodic linking[74] (cf. Torres Cacoullos & Travis 2019).

With respect to ANIMACY, for which the application value was ANIMATE, when the target antecedent bore COLLECTIVE (LOD = 1.73; $p < 0.01$) or INANIMATE (LOD = 1.63; $p < 0.001$)

[74] Contexts in which the clause containing the target subject is BOTH prosodically and syntactically linked to the adjacent prior clause containing its target did not achieve significance in this model, but do achieve significance in the mixed-effects binomial model, Model 3, Section 6.3, ahead.

reference, this was associated with a favoring effect on the realization of a Ø target over a SC. Antecedent subjects with INANIMATE referents exerted an opposite, disfavoring effect on the realization of a X2SBJ (LOD = -11.54; $p < 0.001$), as opposed to a SC. This disproportionately strong effect from INANIMATE antecedents can be understood upon examining Figure 15 which displays the jitter plot with the descriptive results for ANIMACY. One can see that X2SBJs never occurred when their antecedent bore INANIMATE reference (and only twice with COLLECTIVE reference). This is the second invariable context identified thus far.

Figure 15. Jitter plot of SPE x ANIMACY, Analysis 1, broadest envelope.

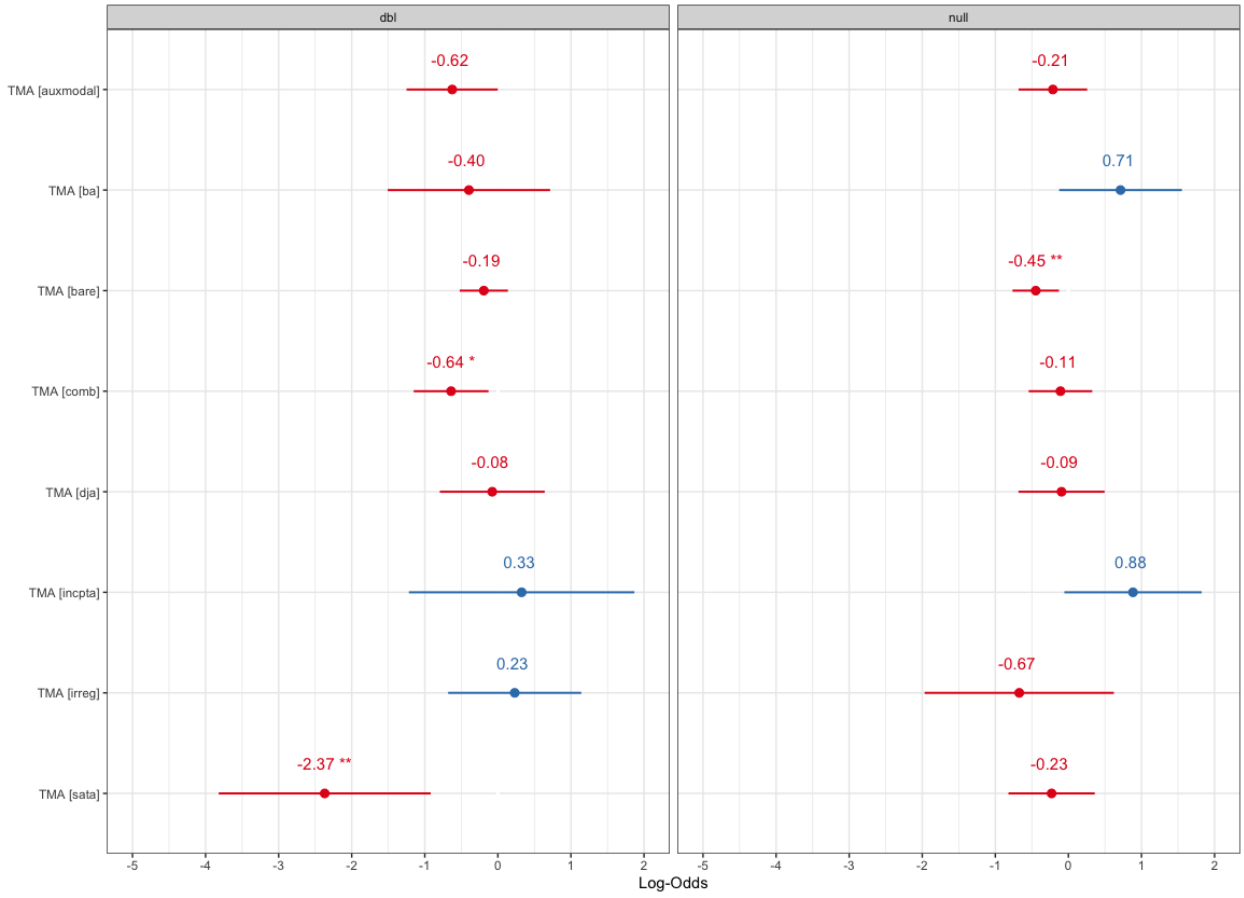


The factor PERSON/NUMBER (1SG, 1PL, 3SG, 3PL) (Section 5.5.2.2) achieved significance for 3PL targets. When compared to 1SG referents, 3PL target subjects favored Ø (LOD = 0.71; $p < 0.01$), but disfavored X2SBJs (LOD = -0.91; $p < 0.05$), as opposed to SCs, respectively. These results

are due in part to INANIMATE and INDEFINITE referents necessarily being 3rd person, though as we will see, the 3rd person favoring effect on Ø subjects cannot solely be attributed to these semantic properties of the referent.

The next factor examined was the overt TMA marking (predictor levels: BARE, AUX/MODAL, *TA*, *-BA*, *SATA*, *DJA*, INCORPORATED *TA*, IRREGULAR, COMBINATION) (Section 5.5.2.3) in the verb phrase for which the target anaphoric subject was an argument (Figure 16). The application value for this level was the highly productive, multi-purpose TMA marker *TA*.

Figure 16. The effects of TMA (application value = *TA*) on SPE, MLR 1, Analysis 1, broadest envelope.

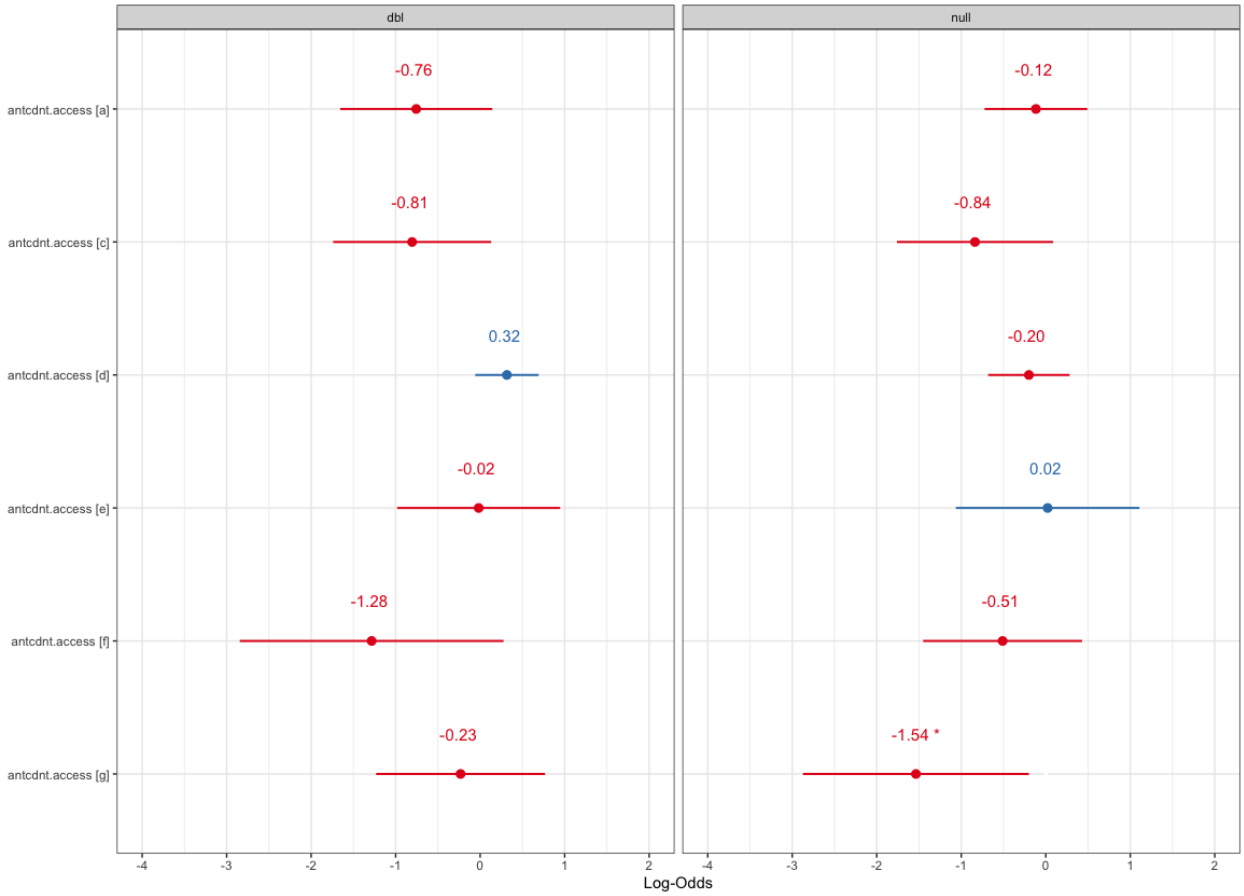


When the VP contained progressive marker *SATA*, X2SBJs were disfavored (LOD = -2.37; $p < 0.01$), when compared to a SC. When the VP contained a COMBINATION of TMA markers and modal/auxiliaries, or several TMA markers simultaneously, this was another disfavoring context

for X2SBJs (LOD = -0.64; $p < 0.05$), as opposed to SCs. BARE verbs were the only level to achieve significance with respect to Ø targets; BARE verbs disfavored the realization of Ø (LOD = -0.45; $p < 0.01$), when compared to a SC. Wagner (2016) found that increased VP complexity exerted a favoring effect on 1st person Ø subjects in English (Section 4.5.3). This could be why Ø is disfavored with BARE verbs (low VP complexity) and why X2SBJs (equivalent in value to singleton tonic pronouns in languages lacking X2SBJ constructions) are disfavored when complexity is introduced into the VP in the form of TMA particles, model/auxiliaries, or some combination of these.

The next most important factor according to the variable importance plot was SESSCORE. However, this will be presented along with the other language-external predictors. For now, I will present the results for ANTECEDENT ACCESSIBILITY PATTERN (predictor levels: PATTERN A, PATTERN B, PATTERN C, PATTERN D, PATTERN E, PATTERN F, PATTERN G) (Section 5.5.2.1). The application value for this predictor was PATTERN B, which referred to contexts in which the antecedent was the subject of the adjacent immediately prior clause (Figure 17).

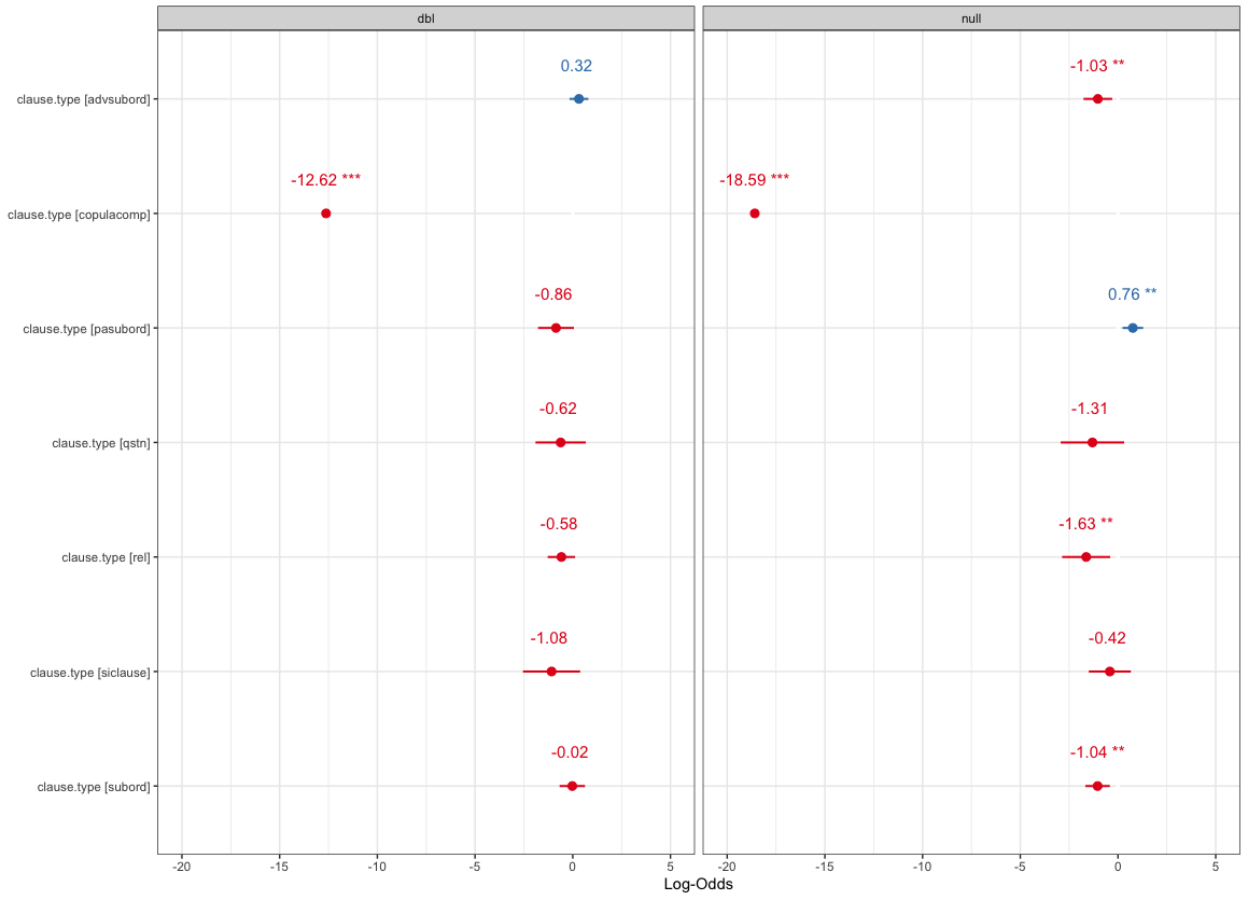
Figure 17. The effect of ANTECEDENT ACCESSIBILITY PATTERN (application value = PATTERN B) on SPE, MLR 1, Analysis 1, broadest envelope.



Only PATTERN G, in which the antecedent was contained in a separate 'discourse chunk' from that of the anaphor, was returned significant. When compared to PATTERN B, an antecedent-target configuration like PATTERN G was disassociated with the realization of target Ø (LOD = -1.54; $p < 0.05$), as compared to a SC. This is to be expected, since Ø anaphora are consistently disfavored in switch-reference contexts (cf. Otheguy & Zentella 2012; Carvalho, Orozco, & Lapidus Shin eds. 2015; Duarte & Soares da Silva 2016) and generally do not reintroduce old, anaphorically distant discourse antecedents since these are low in referential accessibility (cf. Givón 1983a, 2017:6-7; Ariel 1990) (Section 4.5.3).

After ANTECEDENT ACCESSIBILITY PATTERN, the next most important factor was CLAUSE TYPE (predictor levels: MAIN, SUBORDINATE, *PA*-SUBORDINATE, *SI*-CLAUSE, ADVERBIAL SUBORDINATE, RELATIVE, QUESTION, COPULAR COMPLEMENT, COORDINATE) (Section 5.5.2.1), for which the application value was MAIN CLAUSE (Figure 18).

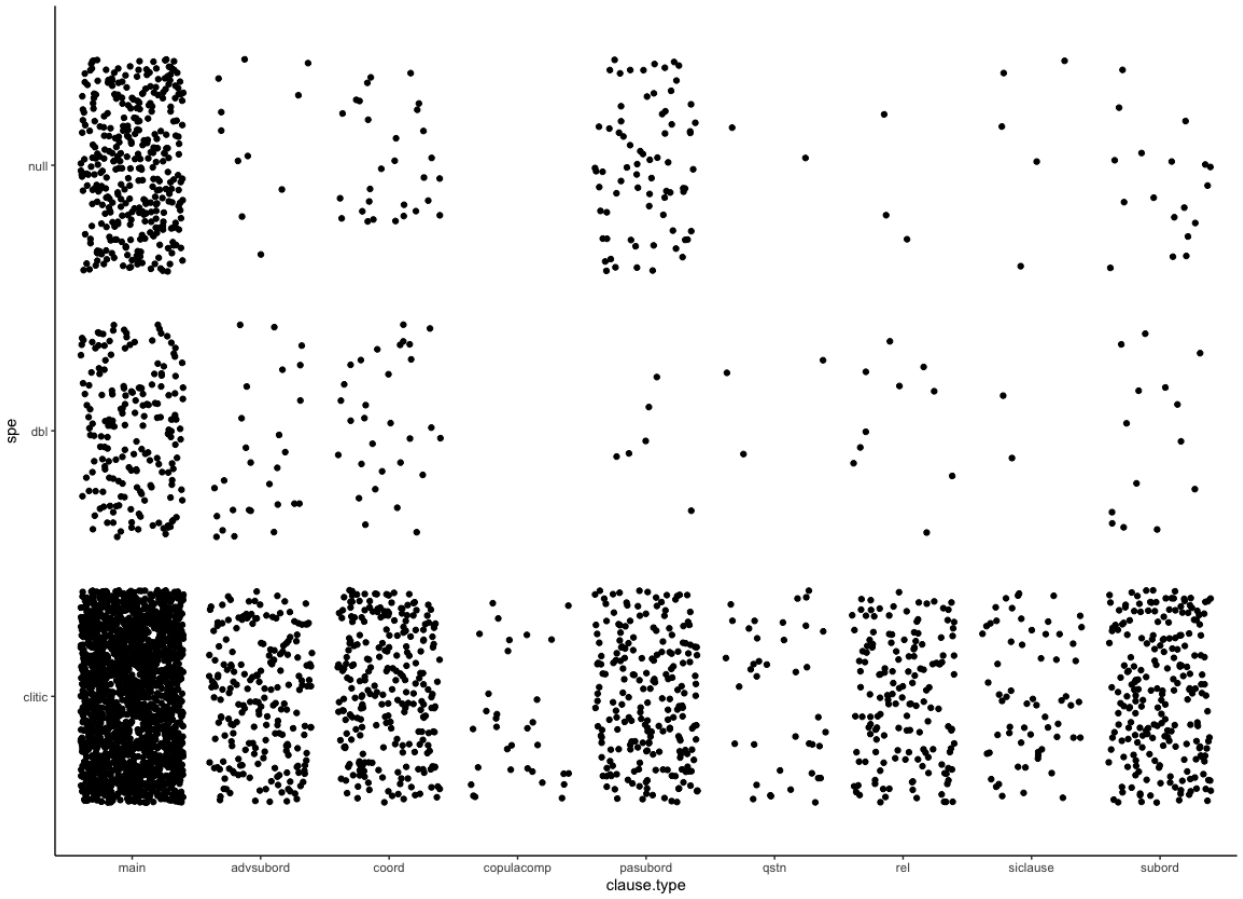
Figure 18. The effect of CLAUSE TYPE (application value = MAIN) on SPE, MLR 1, Analysis 1, broadest envelope.



ADVERBIAL SUBORDINATE clauses were found to disfavor the realization of Ø (LOD = -1.03; $p < 0.01$), as did RELATIVE CLAUSES (LOD = -1.63; $p < 0.01$), and other SUBORDINATE clauses (with complementizers *ki* and *ma*) (LOD = -1.04; $p < 0.01$), when compared to a SC, respectively. This indicates a generalized disfavoring effect on embedded Ø. Both Ø targets (LOD = -18.59; $p < 0.001$) and X2SBJS (LOD = -12.62; $p < 0.001$) were strongly disfavored in COPULAR COMPLEMENT

clauses. Inspection of the jitter plot in Figure 19 reveals that neither of these subject forms ever occurred in COPULAR COMPLEMENT clauses.

Figure 19. Jitter plot of SPE by CLAUSE TYPE, Analysis 1, broadest envelope.

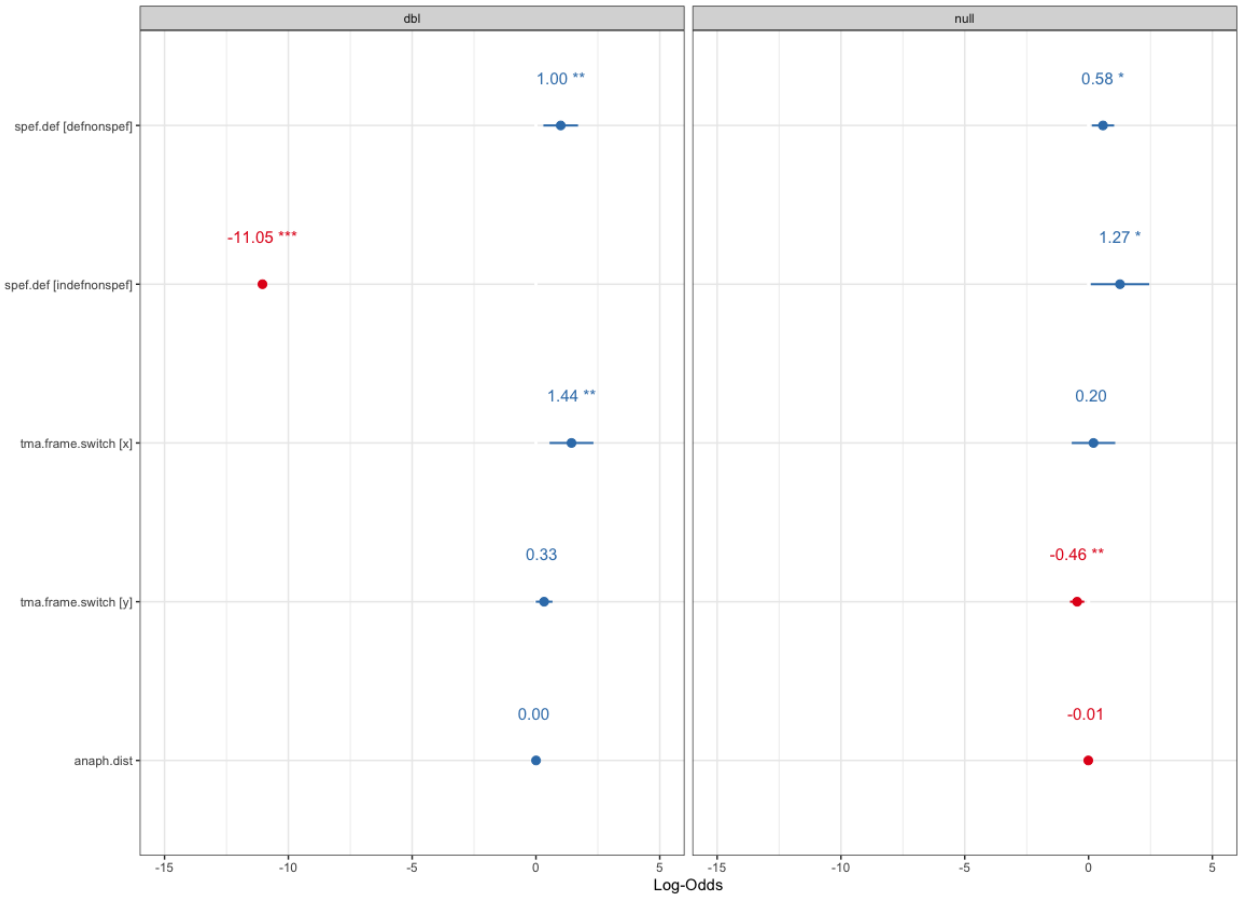


The final three predictors presented (Figure 20) are SPECIFICITY/DEFINITENESS (predictor levels = [+SPECIFIC, +DEFINITE], [+SPECIFIC, -DEFINITE], [-SPECIFIC, +DEFINITE], [-SPECIFIC, -DEFINITE]) (Section 5.5.2.2), TMA FRAME SWITCH (predictor levels = YES, NO, X[75]) (Section 5.5.2.2.3), and ANAPHORIC DISTANCE (continuous numeric) (Section 5.5.2.2). Although SPECIFICITY/DEFINITENESS was the most important of these factors, it will be discussed following the other two, since after presenting the results for SPECIFICITY/DEFINITENESS we will return to our

[75] The level 'X' was when the antecedent was not the subject of the prior clause and was thus diassociated from the TMA 'frame' imposed by the verb on its nominative arguments (as this is relevant for referential continuity/antecedent access).

discussion of the interactions between THE SURFACE FORM OF THE ANTECEDENT, ANAPHORIC DISTANCE, and the semantic properties of the antecedent's subject referent .

Figure 20. The effects of SPECIFICITY/DEFINITENESS (application value = [+SPECIFIC, +DEFINITE]), TMA FRAME SWITCH (application value = NO SWITCH), and ANAPHORIC DISTANCE (continuous numeric) on SPE, MLR 1, Analysis 1, broadest envelope.



The predictor TMA FRAME SWITCH had the application value (N = NO SWITCH). In a VP that experienced a switch in TMA FRAME from the VP for which the antecedent was the subject (Y = SWITCH), Ø targets were disfavored (LOD = -0.46; $p < 0.01$), when compared to SC targets. The disfavoring effect of TMA switches on the realization of Ø anaphora is likely due the increased accessibility burden and decreased referential continuity associated with switching TMA frames from one clause to the next.

Target anaphora with antecedents that were not the subject of their containing clauses (=X), as opposed to NO SWITCH contexts, exerted a favoring effect on the realization of X2SBJs (LOD = 1.44; $p < 0.01$), over SCs. This supports findings for SPE in Spanish that showed that TMA frame switches disfavor Ø (Bayley & Pease-Álvarez, 1997; Geeslin & Gudmestad, 2011) (Section 4.5.3). The favoring effect on X2SBJs when the antecedent was not the subject of its containing clause (and thus not associated with the TMA frame of the VP in that clause) give further indication that X2SBJs have a comparable status to singleton TONIC pronouns: they are used in switch-reference contexts and situations of low antecedent accessibility/referential continuity.

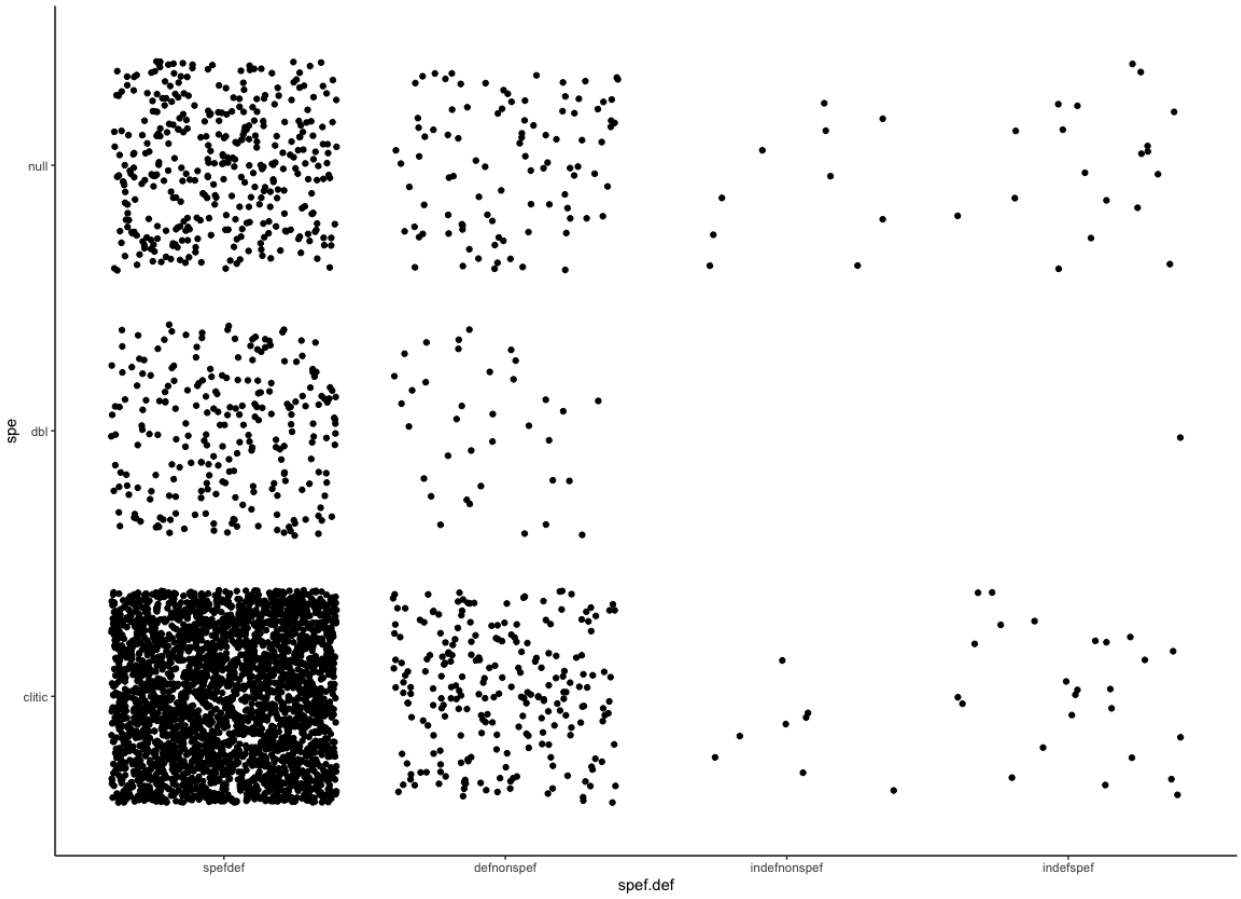
With respect to SPECIFICITY/DEFINITENESS, [+DEFINITE, -SPECIFIC] referents, as opposed to [+DEFINITE, +SPECIFIC] ones, exerted a favoring effect on the realization of a X2SBJ (LOD = 1.00; $p < 0.01$), as well as on the realization of Ø subjects (LOD = 0.58, $p < 0.01$), over a SC, respectively. The former favoring effect was due to contexts in which nonspecific referents were used contrastively. In the following example (4, Section 1.1; repeated here as 144), the antecedent referent for *Anos nu* is *gentis di Maiu* 'people from Maio', while the antecedent referent for *Aes es* is *gentis di praia* 'people from Praia'. In this case the referent in nonspecific (generic) since it refers to a broad collective of people rather than uniquely identified individuals. It was admitted into the envelope because its discourse referents were previously introduced by lexical DPs (see Section 5.4).

(144) *Anos=nu ta fla 'ag', Aes=es ta fla 'agu'*
1.SG.STR=1.SG.CL TMA say 'water' 3.SG.STR=3.SG.CL TMA say 'water'
'We say "ag", they say "agu"'

I will return to the favoring effect exerted by [+DEFINITE, +SPECIFIC] antecedents Ø subjects shortly after discussing the next result for X2SBJs by SPECIFICITY/DEFINITENESS. A [-DEFINITE, -SPECIFIC] referent exerted a strong disfavoring effect on the realization of X2SBJs (LOD = -11.05;

$p < 0.001$), when compared to a SC. As can be seen from the descriptive results in the jitter plot in Figure 21 a X2SBJ only occurred once with an antecedent bearing indefinite reference.

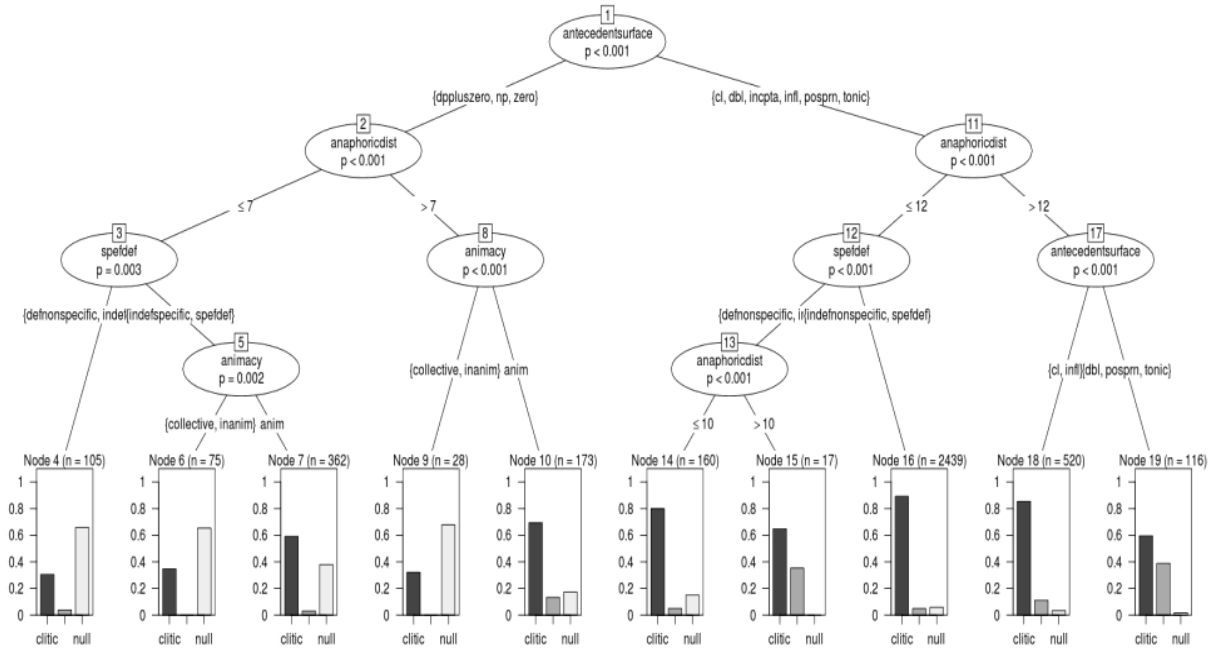
Figure 21. Jitter plot of SPE x SPECIFICITY/DEFINITENESS, Analysis 1, broadest envelope.



This result is consistent with the Typology of Structural Deficiency (Cardinaletti & Starke 1994, 1996, 1999) under which indefinite referents are associated with semantic-referential deficiency, and is thus disassociated from semantic values like indefiniteness. For this reason, X2SBJs almost never bear coreference with indefinite antecedents (cf. Sections 4.3, 4.5.1, and 4.5.2). An antecedent with a [-DEFINITE, -SPECIFIC] referent had an opposite, favoring effect on the realization of a Ø subject (LOD = 1.35, $p < 0.05$), over a SC. This is also consistent with the Typology of Structural Deficiency, since *pro* is assumed to be a deficient pronoun (Cardinaletti & Starke 1994:68,89-91) and is associated with referential deficiency and semantic properties like indefiniteness.

Having now addressed the results for ANIMACY and SPECIFICITY/DEFINITENESS, let us return to the effects for LEXICAL DP and DP + INTERVENING MATERIAL + Ø antecedents on the realization of anaphoric Ø. The conditional inference tree in Figure 22 reveals a relationship between the factors SURFACE FORM OF THE ANTECEDENT, ANIMACY, SPECIFICITY/DEFINITENESS, and ANAPHORIC DISTANCE[76].

Figure 22. Conditional inference tree for the effects of surface form of the antecedent, animacy, specificity/definiteness, and anaphoric distance on SPE, Analysis 1.



The first split in the conditional inference tree reveals a significant effect for SURFACE FORM OF THE ANTECEDENT ($p < 0.001$). Focusing on the left-branching side of the tree, one can see that when the antecedent was a LEXICAL DP, DP + INTERVENING MATERIAL + Ø, or Ø (as opposed to all other forms which split branching right) the ANAPHORIC DISTANCE OF 7 became relevant ($p < 0.001$). At ANAPHORIC DISTANCES >7, ANIMACY becomes relevant ($p < 0.001$), with COLLECTIVE and INANIMATE referents resulting in more Ø subjects when compared to targets with ANIMATE antecedent referents. At $\geq$7 words, SPECIFICITY was relevant ($p < 0.01$), NONSPECIFIC antecedent

[76] Note that anaphoric distance was not returned as significant in the regression models.

referents were resumed by Ø targets at high rates, but also allowed for some X2SBJs. For targets with antecedents bearing SPECIFIC reference, ANIMACY again became relevant ($p < 0.01$); COLLECTIVE and INANIMATE antecedents patterned apart from antecedents with ANIMATE reference. COLLECTIVE and INANIMATE antecedents were associated with high rates of Ø targets while disallowing target X2SBJs.

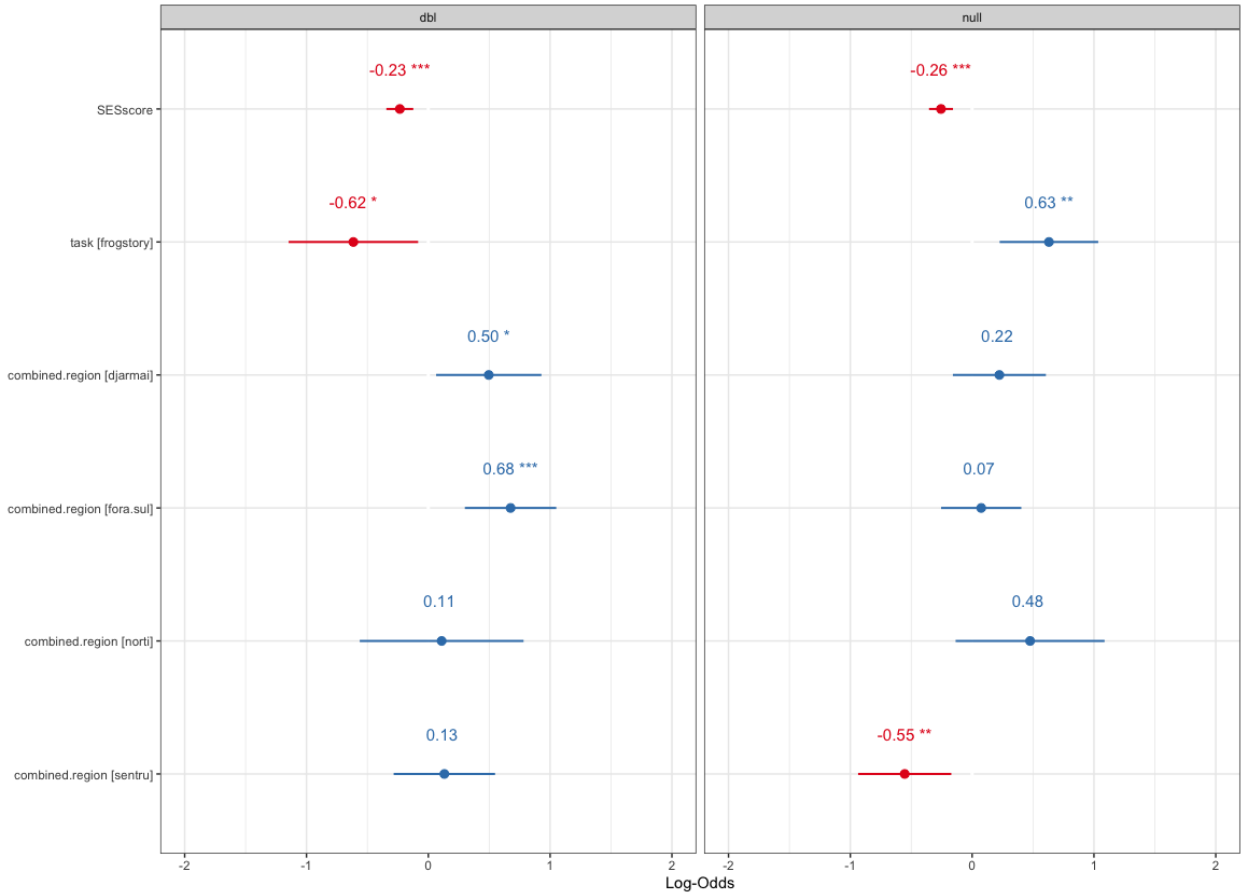
For all other antecedent surface forms, the threshold anaphoric distance was 12 words ($p < 0.001$). When a target lay at an anaphoric distance of >12 words from its antecedent, the SURFACE FORM OF THE ANTECEDENT became relevant ($p < 0.001$). When the antecedent was a TONIC pronoun, X2SBJ, or POSSESSIVE PRONOUN in a LEXICAL DP, this context was associated with higher rates of target X2SBJs. At anaphoric distances of $\leq$12, SPECIFICITY/DEFINITENESS becomes relevant ($p < 0.05$), with antecedents bearing [-SPECIFIC, -DEFINITE] and [+SPECIFIC, +DEFINITE] reference being associated lower rates of both X2SBJs and Ø. For target antecedent with [+DEFINITE, -SPECIFIC] and [-DEFINITE, +SPECIFIC] reference, the anaphoric distance of 10 becomes relevant ($p < 0.01$). In this context, at anaphoric distances >10 (but $\leq$12), there were no instances of Ø, but higher numbers of X2SBJs. At <10 words anaphoric distance, the semantic effects of the antecedent referent begin to exert their effect on the realization of Ø, though the effect is greatly strengthened at $\leq$7 words anaphoric distance. This indicates that the semantic effects of the antecedent's referent on SPE outcomes are operative at short anaphoric distances. The exception might be for antecedents referring to inanimate entities, which pending further empirical confirmation, appear to completely disallow X2SBJ constructions.

Generally speaking, it can be concluded from this set of interactions, that at short anaphoric distances the semantic effects of inanimacy and indefinite reference on lexical DPs conspire to promote Ø. While the effect of animacy ceases to exert a favoring effect on Ø at longer anaphoric distance, its effect on X2SBJs appears to be generalized, since X2SBJs never had inanimate referents associated with their antecedents. Since the promoting effect of the LEXICAL DP and the DP + INTERVENING MATERIAL + Ø construction on Ø was activated when these bore INANIMATE,

INDEFINITE, and NONSPECIFIC reference, these too ceased favoring Ø targets beyond 10 words anaphoric distance. The Ø-to-Ø structural priming effect also operates at short anaphoric distances.

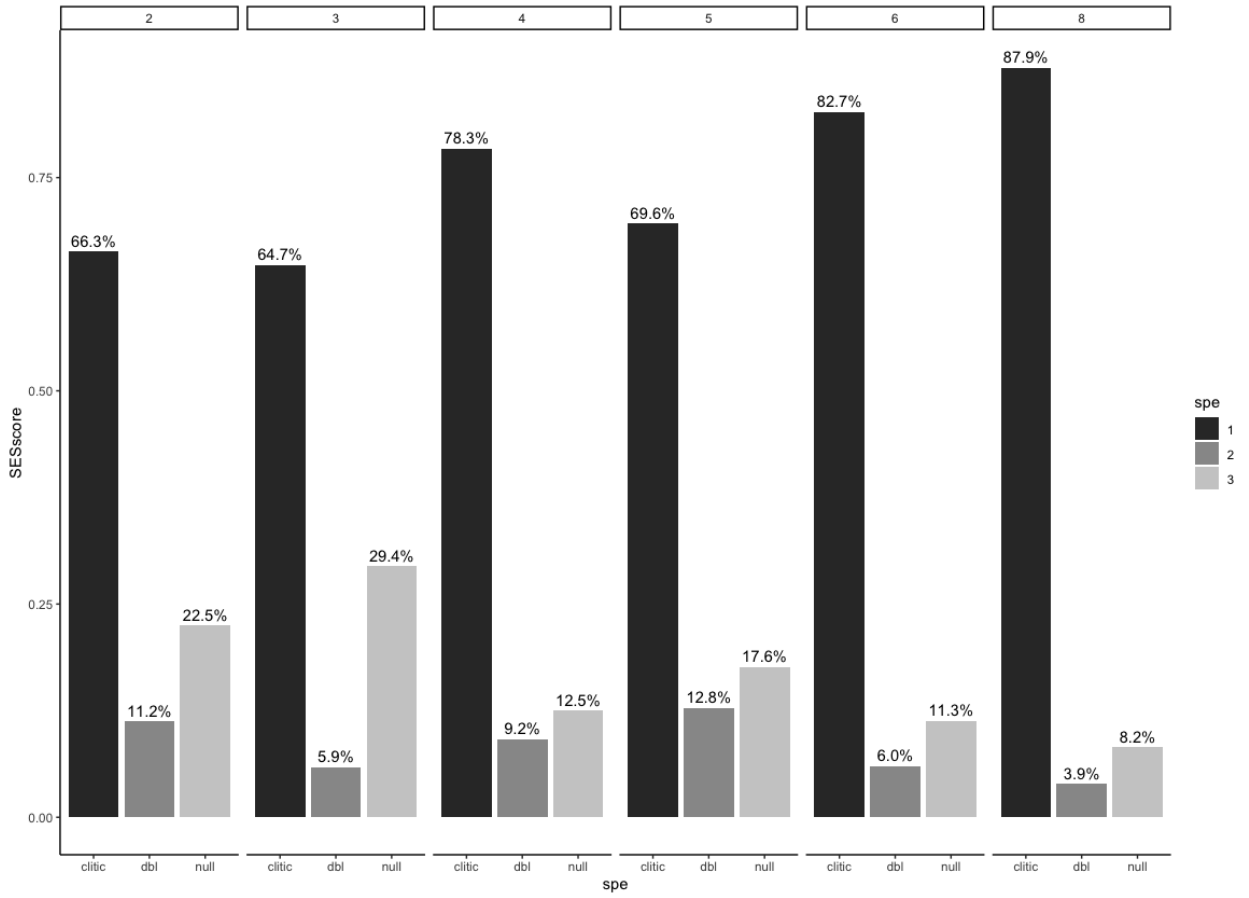
Returning to our discussion of the MLR output, the final set of results for the broad-envelope model are for the four language-external factors (Section 5.5.2.4), SESSCORE (continuous numeric, 2-8), COMBINED DIALECT REGION (predictor levels = SANTIAGO SUL, SANTIAGO NORTI, SANTIAGU SENTRU, DJARMAI, COMBINED), TASK (predictor levels = INTERVIEW, PICTURE DESCRIPTION NARRATIVE), and AGE (continuous numeric, 18-57) (Figure 23). Recalling the variable importance plots in Figures 10 and 11 above, SESSCORE was by far the most important individual-specific factor and fell just below TMA in terms of overall variable importance. TASK was much lower in variable importance ranking, below overt CLAUSE TYPE but before SPECIFICITY/DEFINITENESS.

Figure 23. The effect of SESSCORE (continuous numeric), TASK (application value = INTERVIEW), and COMBINED DIALECT REGION (application value = SUL) on SPE, MLR 1, Analysis 1, broadest envelope.



An increase of magnitude one in a participant's SESSCORE was found to significantly disfavor the realization of both X2SBJs (LOD = -0.23; $p < 0.001$) and Ø (LOD = -0.26; $p < 0.001$), when these outcomes were compared with SCs, respectively. These results indicate that speakers with higher SESSCOREs realized less of a variety of subject forms and relied on more SCs when resolving anaphoric relationships across the discourse. A bar plot of the distribution of subject forms by SESSCORE is shown in Figure 24.

Figure 24. Bar plot of SPE x SESSCORE, Analysis 1, broadest envelope.



The factor TASK, with application value INTERVIEW, was also significant. The PICTURE DESCRIPTION NARRATIVE ('The Frog Story') favored the realization of Ø (LOD = 0.84, $p < 0.001$) and disfavored the realization of X2SBJs (LOD = -0.56, $p < 0.05$), when compared to SCs, respectively. The former outcome is likely related to the predominance of 3rd person referents in the PICTURE DESCRIPTION NARRATIVE, while the latter result is likely due to the continued use of the same three referents (a boy, a dog, and a frog) throughout the narrative, thus reducing the need to mark switch reference and to reintroduce referents after long anaphoric distances. The results may, however, be related to narrative style: story-telling may be associated with different SPE patterns than conversational-style.

The dialect regions for the speakers in the sample were discussed in the previous chapter (Section 5.3.2, 5.5.2.4, and 5.6). It was mentioned that Santiago can be split into roughly three major dialect zones (Quint p.c.) corresponding to the levels SANTIAGU NORTI, SANTIAGU SENTRU,

SANTIAGU SUL (see Image 3, Section 5.6); Maio constituted its own dialect zone, DJARMAI (Section 5.3.2.2). It was also noted that Cape-Verdeans regularly make popular stereotypical distinctions about residents of Praia as opposed to residents who live *fora* 'outside', referring to any area outside of the capital. This distinction appear to at least in part be informed by linguistic perceptions regarding the style, register, or variety used by speakers in the capital. Further, many Cabo-Verdeans migrate to Praia for reasons of work, education, general opportunity, and personal relationships. Five speakers in the sample were originally from *fora* but have lived and worked in Praia in their adulthood; one speaker was from DJARMAI, one was from SANTIAGU NORTI, and three were from SANTIAGU SENTRU (Table 33, Section 5.5.2.4, recast as Table 37).

Table 37. Speakers' primary childhood and adulthood dialect regions, repeated

Dialect zone	Dialect region in childhood	Dialect region in adulthood
SANTIAGO SUL	12	17
SANTIAGO SENTRU	10	7
SANTIAGO NORTI	4	3
DJARMAI	7	6
Total	33	33

The predictor COMBINED DIALECT REGION was thus divided into five groups, four groups were for speakers who resided in the same place they were born and raised, and the other group was for speakers that were from *fora* but who live and reside in Praia in adulthood. Thus, the levels for COMBINED DIALECT REGION were: SANTIAGO SUL, SANTIAGO SENTRU, SANTIAGO NORTI, DJARMAI, and COMBINED (SUL + *FORA*). The results for COMBINED DIALECT REGION show that speakers from DJARMAI (LOD = 0.50; $p < 0.05$), as well as those who grew up *fora* but resided in Praia in their adulthood (COMBINED = *FORA* + SUL) (LOD = 0.68; $p < 0.001$), used significantly more X2SBJs when compared with SCs.

An examination of the variance inflation factor associated with the predictors in Model 1 was conducted using the `vif` command in the `usdm` package (Naimi *et al.* 2017) in R (R Core Team 2019). This revealed extreme multicollinearity in the model. The restricted-envelope models

considered in the following subsections were thus examined for multicollinearity before retaining the analysis for discussion of the results.

In this subsection, several contexts of non-variability were observed in which not all three of the possible subject forms were used. These included a level for the predictor SURFACE FORM OF THE ANTECEDENT for antecedents associated with the putative (medial) subject position of the INCORPORATED *TA* serializing construction (excluded X2SBJs), targets whose antecedent bore INANIMATE reference (excluded X2SBJs) or INDEFINITE reference (nearly categorically exclude X2SBJs), and copular complement clauses (excluded both X2SBJs and Ø). In the next analysis, in order to constrain the envelope of variation based on empirical priors, these contexts were excluded. The next analysis also applies additional restrictions on admissible Ø subjects; the MLR Model 2 in Analysis 2 will exclude target subjects that were the arguments of bare verbs from classes 3 and 4 when these received a nonpast reading. Verbs from these classes have been hypothesized to only be tensed in their bare form when attributed a simple past reading (cf. Silva 1990; see Section 5.4).

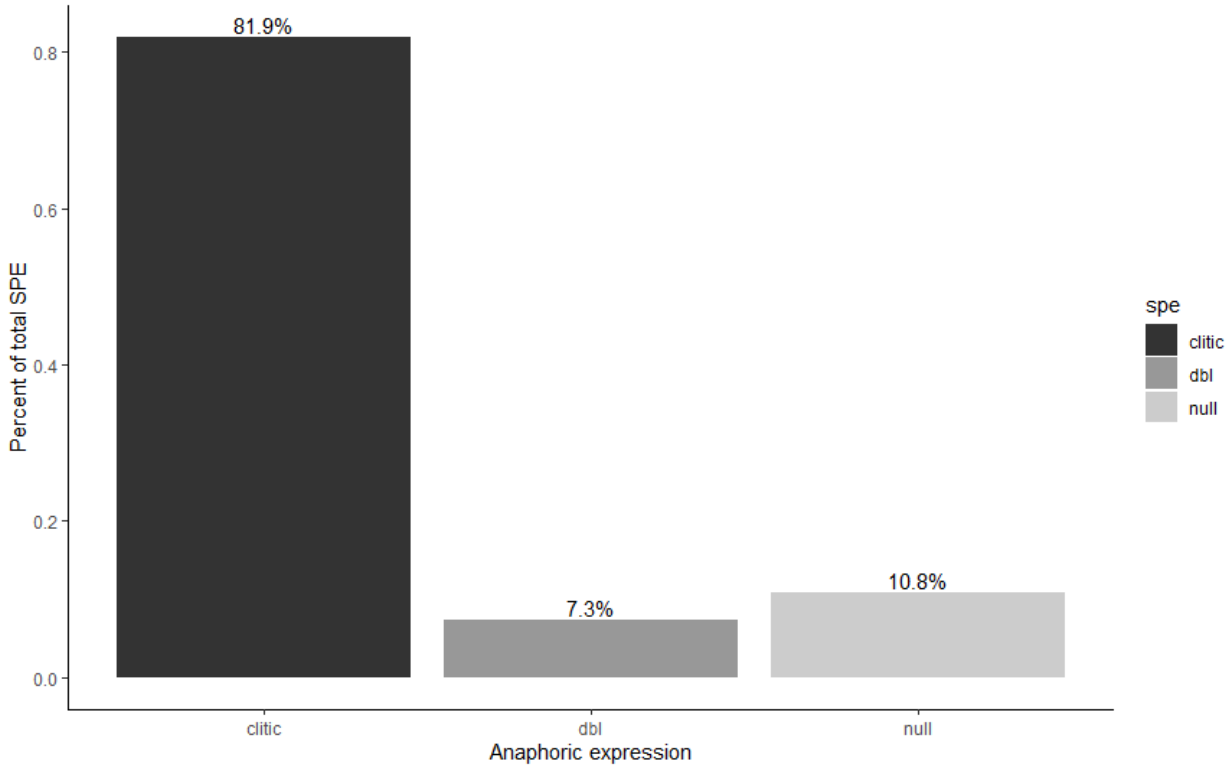
6.2 ANALYSIS 2: MLR 2, CONSTRAINED ENVELOPE OF VARIATION.

The second analysis, like Analysis 1, considered a three-way response variable for SPE (SC, X2SBJ, Ø). Analysis 2 removes several contexts of non-variability that were present in Analysis 1, as well as a contexts in which the referentiality of Ø could be called into question on grounds of dubious finiteness or argumental status. With respect to invariable contexts, we saw that X2SBJs never occurred with inanimate referents and there was only once instance of a X2SBJ with an antecedent bearing an indefinite referent. Therefore, the predictor ANIMACY was removed along with observations involving INANIMATE and COLLECTIVE referents. SPECIFICITY/DEFINITENESS was relabeled SPECIFICITY with levels SPECIFIC and NONSPECIFIC and observations with INDEFINITE reference were removed.

Another site of invariability was with respect to the SURFACE FORM OF THE ANTECEDENT; the putative (medial) subject position of the INCORPORATED *TA* construction was never an antecedent to a target X2SBJ. All observations of target anaphora bearing a INCORPORATED *TA* construction were removed from the model. Another delimitation involved the removal of COPULA COMPLEMENT clauses, since Ø subjects and X2SBJs never occurred in these constructions. Recalling the stativity and controllability-based classification of CVC verbs established in Silva (1990) (Sections 5.4 and 5.5.2.3), VERB CLASSes 3 and 4 are attributed a default past reading when bare, while VERB CLASSes 1 and 2 get a default nonpast reading when bare. There did occur some instances of bare CLASS 3 and 4 verbs that could not be attributed a past reading (cf. example 143; Section 5.4); these were removed, as they might be argued to be not fully finite or their subject to not genuinely be referential or argumental.

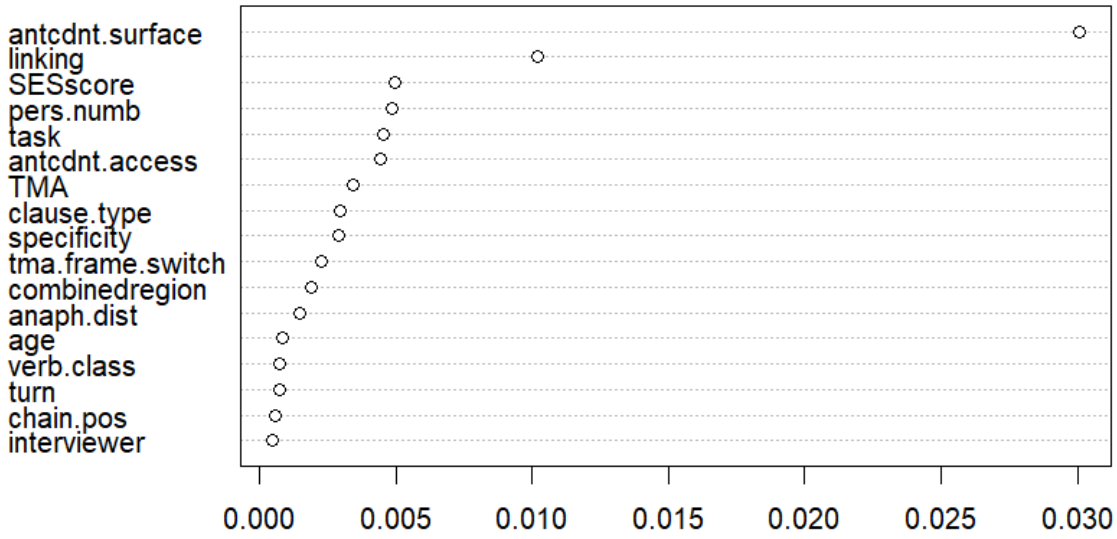
The dataset for Analysis 2 was characterized by this more restricted envelope and contained a total of 3,651 subject realizations. The overall share of subject forms for Ø, SCs, and X2SBJs are shown in the bar plot in Figure 25. The proportion of SCs was roughly the same, 81.9%. Ø remained the second most frequent subject type overall at a rate of 10.8%. The proportion of X2SBJs increased slightly to 7.3%.

Figure 25. Proportion of target subjects, bar plot, Analysis 2, restricted envelope.



A random forest was grown and plotted for variable importance (Figure 26). The random factor PARTICIPANT was excluded from the random forest in Analysis 2, since the MLR for Analysis 2 was a fixed-effects model, but was included in the first variable importance plot in Model 1 (Figure 10, above), and is considered in mixed-effects binomial regression models in Analyses 3 and 4.

Figure 26. Variable importance plot of random forest for Analysis 2, constrained envelope.



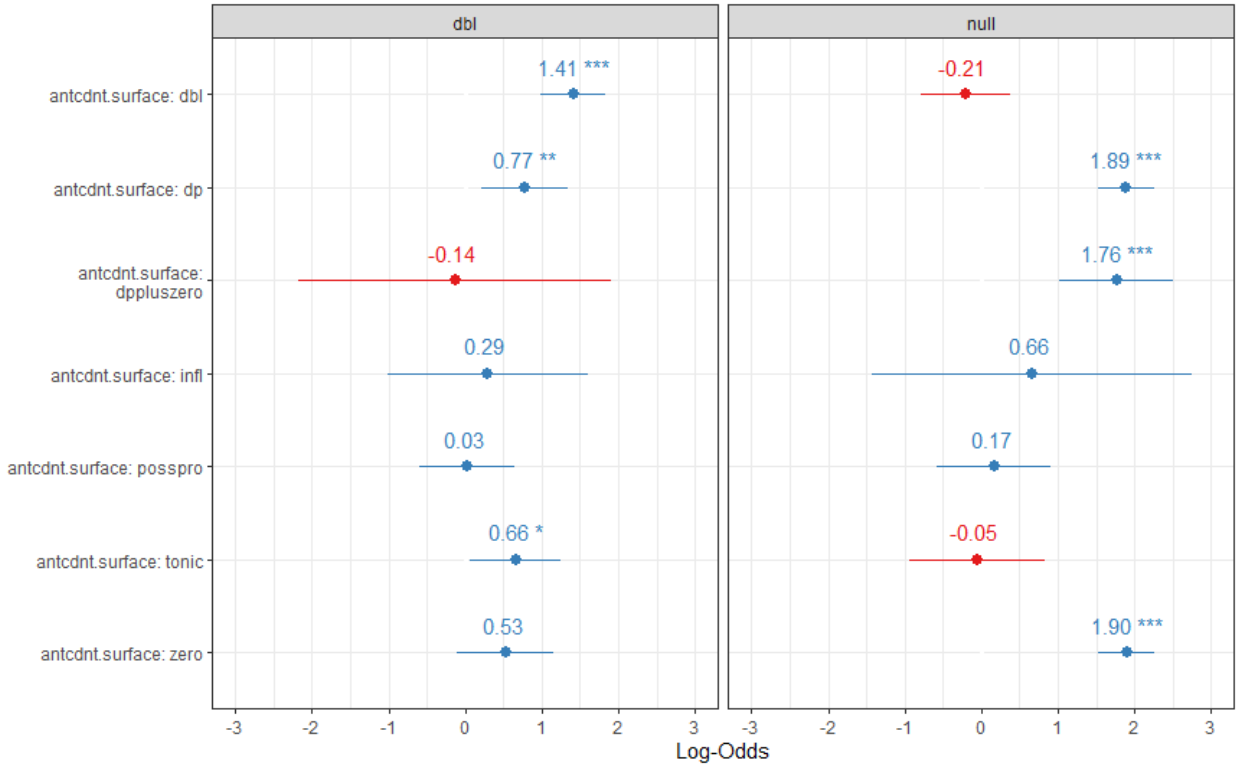
Just as in Analysis 1, the most important predictor was SURFACE FORM OF THE ANTECEDENT, associated with linguistic structural priming, and this factor was again followed by LINKING. SESSCORE rose to the third most important ranking and PERSON/NUMBER moved up to the fourth. Next came TASK, followed by ANTECEDENT ACCESSIBILITY PATTERN, then TMA. After this was CLAUSE TYPE which maintained a comparable ranking to Model 1, as did SPECIFICITY, TMA FRAME SWITCH, COMBINED DIALECT REGION, ANAPHORIC DISTANCE, AGE, and so forth.

Just as with the previous analysis, variables were entered into the regression as model parameters following their order of variable importance (from the random forest), and then each parameter was eliminated in a backwards selection procedure that resulted in a 'base' model. The 'base' model informed the design of several other models with comparable set of parameters, but with different interaction terms or single parameters removed. These were models were then nested and compared for AIC scores along with the first-to-converge 'base' model using the `step`

function. The model[77] of best fit included the predictors SURFACE FORM OF THE ANTECEDENT, LINKING, SESSCORE, ANTECEDENT ACCESS PATTERN, PERSON/NUMBER, TASK, TMA, CLAUSE TYPE, SPECIFICITY, TMA FRAME SWITCH, ANAPHORIC DISTANCE, COMBINED DIALECT REGION, and AGE.

The results for SURFACE FORM OF THE ANTECEDENT[78] (Section 5.5.2.1) in Model 2 are shown in Figure 27. There were no changes from MLR 1 in Analysis 1 in the predictor levels that achieved significance.

Figure 27. The effect of SURFACE FORM OF THE ANTECEDENT (application value = CLITIC) on SPE in MLR 2, Analysis 2, constrained envelope.



A strong priming effect remained; a X2SBJ antecedent, as opposed to a CLITIC, exerted a favoring effect on the realization of a X2SBJ target (LOD = 1.41; $p < 0.001$), as did TONIC (monosyllabic and disyllabic; oblique, subject or object) pronouns (LOD = 0.66; $p < 0.05$), and

[77] Residual Deviance: 3336.629; AIC: 3520.629

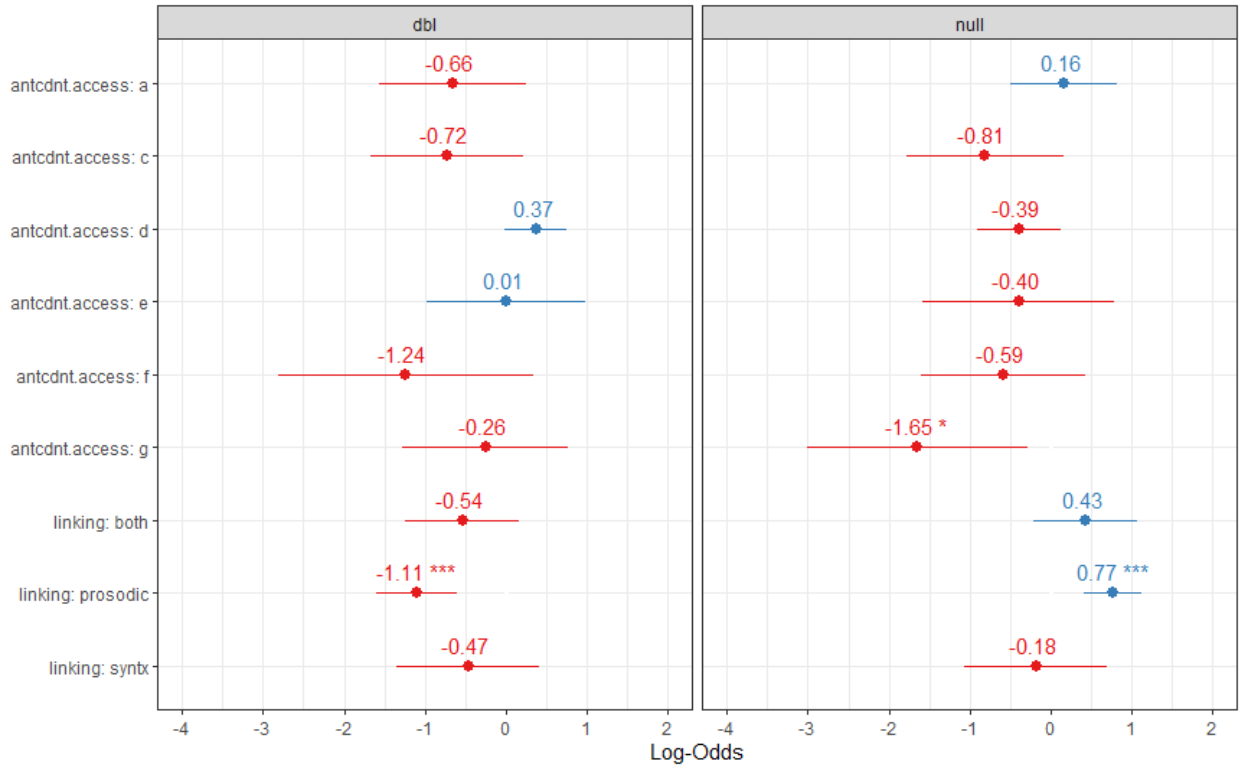
[78] (predictor levels = X2SBJ, LEXICAL DP, DP + INTERVENING MATERIAL + Ø, INFLECTION, POSSESSIVE PRONOUN, TONIC, Ø),

LEXICAL DPs (LOD = 0.77; $p < 0.05$), when these were compared to a SC, respectively. This last result, for which LEXICAL DPs antecedents exerted a favoring effect on X2SBJs, is not a priming effect; rather, these were all switch-reference contexts in which the antecedent to the X2SBJ was in some configuration other than that of the subject of the prior clause (ANTECEDENT ACCESSIBILITY PATTERNS C, D, E, D, or F). There were only three of twenty-nine (=10%) observations of a LEXICAL DP as an antecedent to a X2SBJ that were not switch-reference contexts. These results for SURFACE FORM OF THE ANTECEDENT reinforce the observed X2SSBJ-to-X2SBJ priming effect. It also reinforces the notion that X2SBJs have an equivalent priming value to singleton TONIC pronouns and that X2SBJs are used as switch-reference devices, particularly when contrastively opposed to other discursively prominent referring expression like lexical DPs.

The structural priming effect for Ø antecedents remained as well; Ø antecedents, as opposed to CLITICs, were predictive of Ø target subjects (LOD = 1.90; $p < 0.001$), over SCs. The effects of the antecedent that were not attributable to priming also held. A LEXICAL DP antecedent, as opposed to a CLITIC, exerted a favoring effect on target Ø subjects (LOD = 1.89; $p < 0.001$), over SCs. The promoting effect of a DP + INTERVENING MATERIAL + Ø construction antecedent, as opposed to a CLITIC, remained; these antecedents promoted anaphoric Ø (LOD = 1.76; $p < 0.001$), over SCs. I will once again examine a conditional inference tree to inspect why LEXICAL DPs and the DP + INTERVENING MATERIAL + Ø construction promoted Ø subjects. This last set of non-priming related results was again associated with the semantic effects of SPECIFICITY (indefinite and inanimate referents having been removed). I will return to these results later in this subsection when presenting the results for SPECIFICITY.

The next two significant factors were LINKING (predictor levels = NO LINK, PROSODIC, SYNTACTIC, BOTH) (Sections 4.5.3, 5.5, and 5.5.2.1) and ANTECEDENT ACCESSIBILITY PATTERN (predictor levels = PATTERNS A, B, C, D, E, F, G) (Sections 4.5.1 and 5.5.2.1). Some different outcomes from MLR1 in Analysis 1 to MLR 2 in Analysis 2 obtained for the latter predictor (Figure 28).

Figure 28. The effects of LINKING (application value = NO LINK) and ANTECEDENT ACCESSIBILITY PATTERN (application value = PATTERN B) on SPE, MLR 2, Analysis 2, constrained envelope.



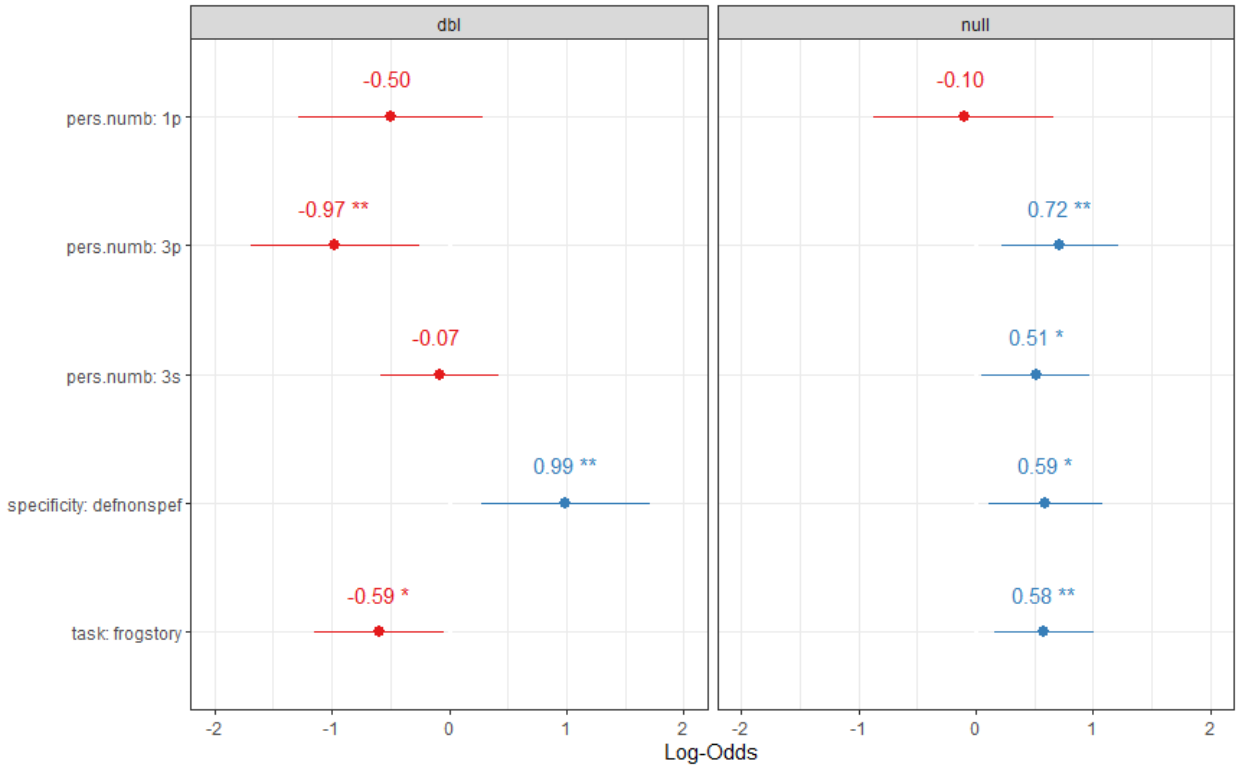
For LINKING, once again, PROSODIC LINKING, as opposed to NO LINKING, was found to disfavor the realization of a X2SBJ (LOD = -1.11; $p < 0.01$), but favored the realization of anaphoric Ø (LOD = 0.77; $p < 0.01$), when these were compared with SCs, respectively. These results reiterate the importance of PROSODIC LINKING in promoting Ø subjects cross-linguistically (cf. Torres Cacoullos & Travis 2019).

The next most important factor that achieved significance in the model was SESSCORE, but this predictor will be addressed along with the other language-external individual-specific factors at the end of this subsection. For the time being, I skip ahead to the next most important factor, ANTECEDENT ACCESSIBILITY PATTERN. PATTERN G contexts, in which the antecedent was contained in a 'discourse chunk' distinct from that of the target, remained a disfavoring context for the

realization of a Ø subjects (LOD = 1.65; $p < 0.05$), when compared to a SC. These results support the notion that X2SBJs are a "switch-reference device" while Ø is a "high referential continuity device" (Givón 1983b; 2017:5-8) and that X2SBJs generally exercise the same role as singleton tonic pronouns in terms of discourse organization and referential continuity; primarily, they serve to reintroduce old discourse referents.

The next most important factors were PERSON/NUMBER (predictor levels = 1SG, 1PL, 3SG, 3PL) (Section 5.5.2.2) and TASK (predictor levels = INTERVIEW, PICTURE DESCRIPTION NARRATIVE) (Section 5.5.2.4). These are presented alongside a less important factor (Figure 33), SPECIFICITY (predictor levels: SPECIFIC, NONSPECIFIC) (Figure 29), since as we will see in turn, these predictors interacted in important ways.

Figure 29. The effects of PERSON/NUMBER (application value = 1SG), SPECIFICITY (application value = SPECIFIC), and TASK (application value = INTERVIEW) on SPE, MLR 2, Analysis 2, constrained envelope.



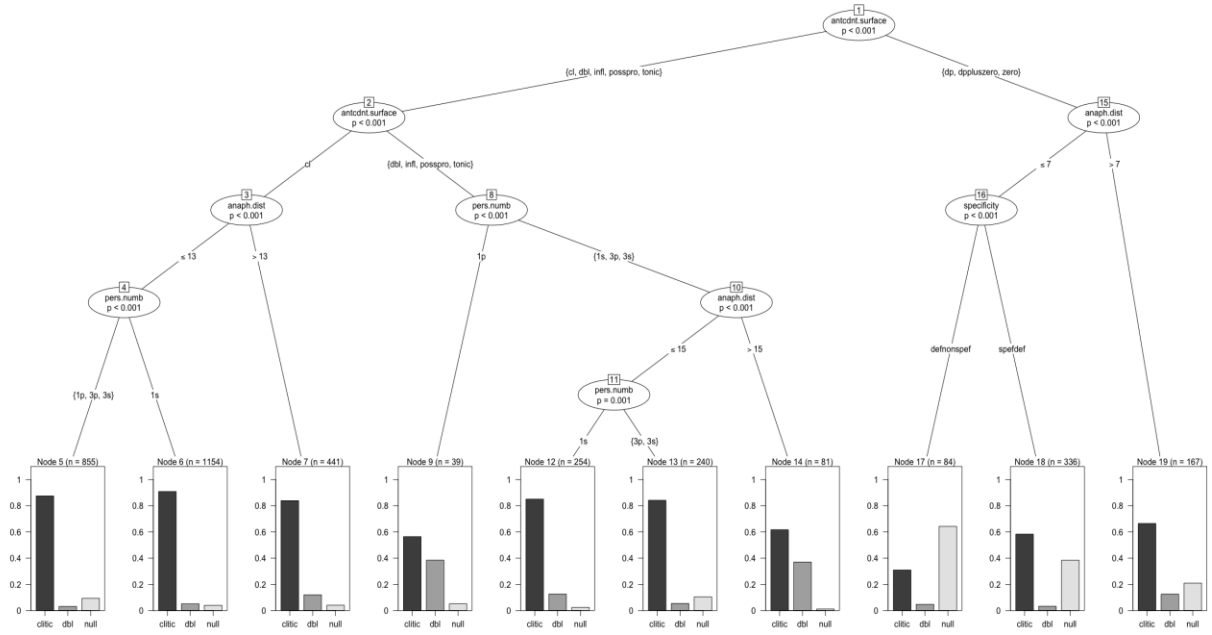
The results for PERSON/NUMBER in MLR 2 remain mostly unaltered from MLR 1, although the overall predictor-level significance and the magnitude of their effect is greater in MLR 2. 3PL referents, as opposed to 1SG ones, disfavored the realization of a X2SBJ (LOD = -0.97; $p < 0.01$), but favored the realization of Ø (LOD = 0.72; $p < 0.01$), when these were compared with SCs, respectively. 3SG referents, as opposed to 1SG ones, also favored the realization of Ø subjects (LOD = 0.59; $p < 0.05$), when these were compared with SCs, respectively.

The effects for SPECIFICITY of the antecedent referent also held in MLR 2, even following the removal of the observations associated with DEFINITENESS for this predictor. As before, when the semantic reference of the antecedent was [-SPECIFIC], as opposed to [+SPECIFIC], there was a favoring effect on the realization of a X2SBJ (LOD = 0.99; $p < 0.01$), as well as on the realization

of Ø (LOD = 0.61, $p < 0.05$), over a SC, respectively. The results for X2SBJs can be attributed to their use with nonspecific referents in cases of switch-reference and contrast (example 144, Section 6.1).

As in Analysis 1, an inspection of a conditional inference tree for predictors SURFACE FORM OF THE ANTECEDENT, ANAPHORIC DISTANCE, SPECIFICITY and PERSON/NUMBER (Figure 30) revealed that the effects for SPECIFICITY mostly obtain at short anaphoric distances, with nonspecific LEXICAL DPs and DP + INTERVENING MATERIAL + Ø antecedents promoting Ø target subjects at $\leq$7 words distance. Thus, the relationship between semantically referentially deficient lexical DPs (and similar antecedents) and anaphoric Ø is unlocked at short anaphoric distances. This findings lends credence to the notion that the same "avoid referential deficient pronoun" constraint is probabilistically active in CVC, just like in Brazilian Portuguese (cf. Kato & Duarte 2002, 2005; Duarte & Soares da Silva 2016) (Section 4.5.1).

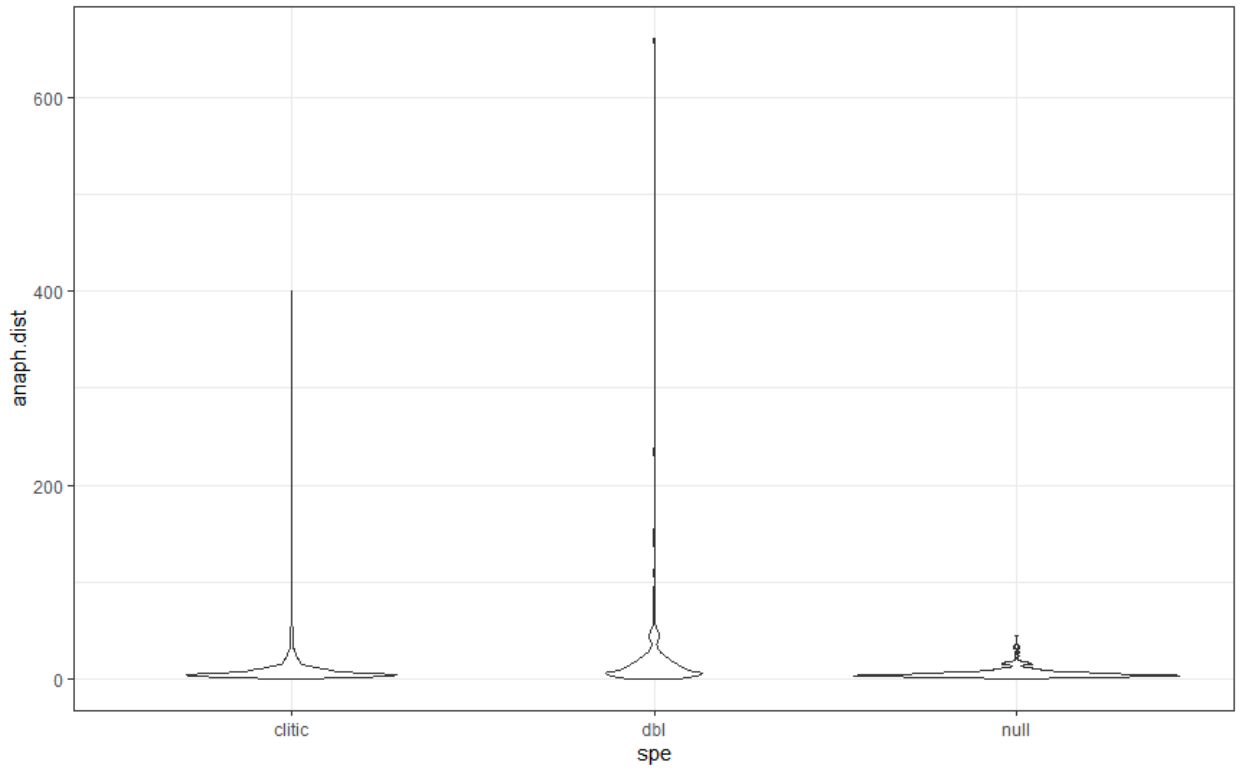
Figure 30. Conditional inference tree for SURFACE FORM OF THE ANTECEDENT, ANAPHORIC DISTANCE, SPECIFICITY and PERSON/NUMBER, Analysis 2, constrained envelope.



The promoting effect of LEXICAL DPs and DP + INTERVENING MATERIAL + Ø antecedents on X2SBJs becomes relevant at >7 words distance, indicating that X2SBJs are being used contrastively in opposition to referents recently introduced by lexical DPs. The Ø-to-Ø target priming effect is also strongest at ≤7 words distance. Figure 31 displays a violin plot for the effect of ANAPHORIC DISTANCE on SPE[79]. The width of the violin figure represents the number of tokens occurring at a given anaphoric distance (y-axis = number of words), while the height of the plot represents the maximum anaphoric distance that occurred with each subject type (x-axis). It can be observed that X2SBJs occurred at long anaphoric distances, anaphoric Ø clusters at short anaphoric distances, and SCs mostly cluster at short distances but extend across a greater range than Ø subjects.

[79] Note that ANAPHORIC DISTANCE, as in MLR 1, did not achieve significance in MLR 2.

Figure 31. Violin plot for the effects of ANAPHORIC DISTANCE on SPE, clitic (SC), double 'dbl' (X2SBJ), and null (Ø), Analysis 2, constrained envelope.



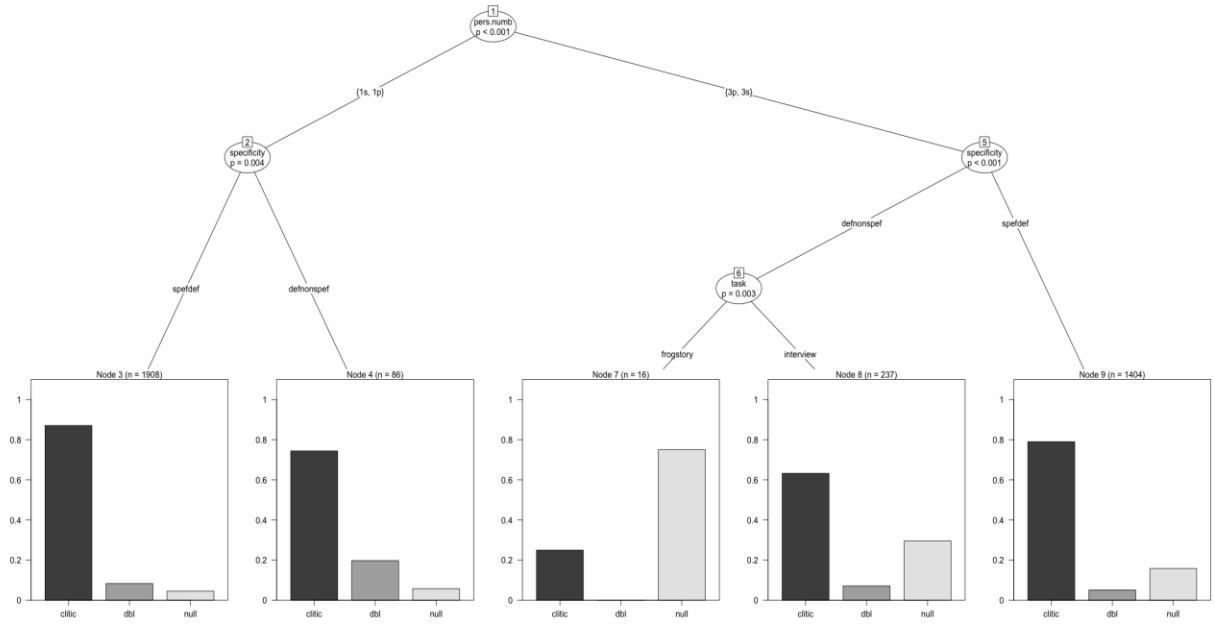
Returning to the conditional inference tree in Figure #, for all antecedent forms other than LEXICAL DPS, DP + INTERVENING MATERIAL + Ø, and Ø, antecedent SPECIFICITY was no longer relevant; instead, PERSON/NUMBER became relevant and interacted with ANAPHORIC DISTANCE. When the antecedent was a CLITIC, 13 words ANAPHORIC DISTANCE was relevant ($p < 0.001$); at $\leq$13 words, target subjects with 1PL, 3SG, and 3PL referents were distinguished from 1SG referents in selecting for more instances of anaphoric Ø ($p < 0.001$). At >13 words, PERSON/NUMBER lost its relevance and X2SBJS occurred more often. When the antecedent was a X2SBJ, a TONIC pronoun, a POSSESSIVE PRONOUN, or superstrate INFLECTION, these were distinguished from CLITICS ($p < 0.001$). PERSON/NUMBER was also relevant for these antecedents ($p < 0.001$), 1PL referents were distinguished from all others in selecting for higher rates of X2SBJS.

For 1SG, 3SG, and 3PL anaphora, an ANAPHORIC DISTANCE of 15 words from the antecedent was relevant ($p < 0.001$). At >15 words X2SBJ, TONIC, POSSESSIVE PRONOUN, and INFLECTION antecedents triggered high rates of X2SBJs. At $\leq$15 words distance, PERSON/NUMBER was again relevant ($p < 0.01$); at this distance, 3rd person referents allowed for more Ø and less X2SBJs. Conversely, at this same distance, more X2SBJs were observed in the 1SG.

These results confirm those hinted at in exploratory Analysis 1: short anaphoric distances cause Ø antecedents to prime Ø target antecedents, and the relationship between Ø anaphora and referentially deficient (i.e. NONSPECIFIC) LEXICAL DPs and DP + INTERVENING MATERIAL + Ø antecedents is also most active at short anaphoric distances. X2SBJs are selected more often at longer distances, and the longer the distance, the more likely X2SBJs are to be associated with 1st person reference.

The results for TASK in Figure # above revealed that the PICTURE DESCRIPTION NARRATIVE was a favoring condition for the realization of a Ø subject (LOD = 0.58; $p < 0.05$), and disfavoring condition for the realization of a X2SBJ (LOD = 0.59; $p < 0.05$), over a SC, respectively. Upon inspection of the interaction between PERSON/NUMBER, SPECIFICITY, and TASK (Figure 32), it was revealed that nonspecific 3rd person referents were responsible for large numbers of anaphoric Ø, particularly during the PICTURE DESCRIPTION NARRATIVE.

Figure 32. Conditional inference tree for the effects of PERSON/NUMBER, SPECIFICITY, and TASK on SPE, Analysis 2, constrained envelope.



This effect was attributable to instances in which participants provided some tangential narrative in which the primary referents were not the three main characters in the 'Frog Story' (the boy, the dog, or the frog) (see Sections 5.5 and 5.5.2.4). For instance, while completing the PICTURE DESCRIPTION NARRATIVE, speaker P16 began describing the activities of typical Cabo-Verdean children playing with *ganfanhotus* 'crickets' during the *azagua* 'rainy season'. This led speakers like P16 to use many nonspecific (generic-like) referents introduced by a lexical DP heading an anaphoric Ø-chain (145).

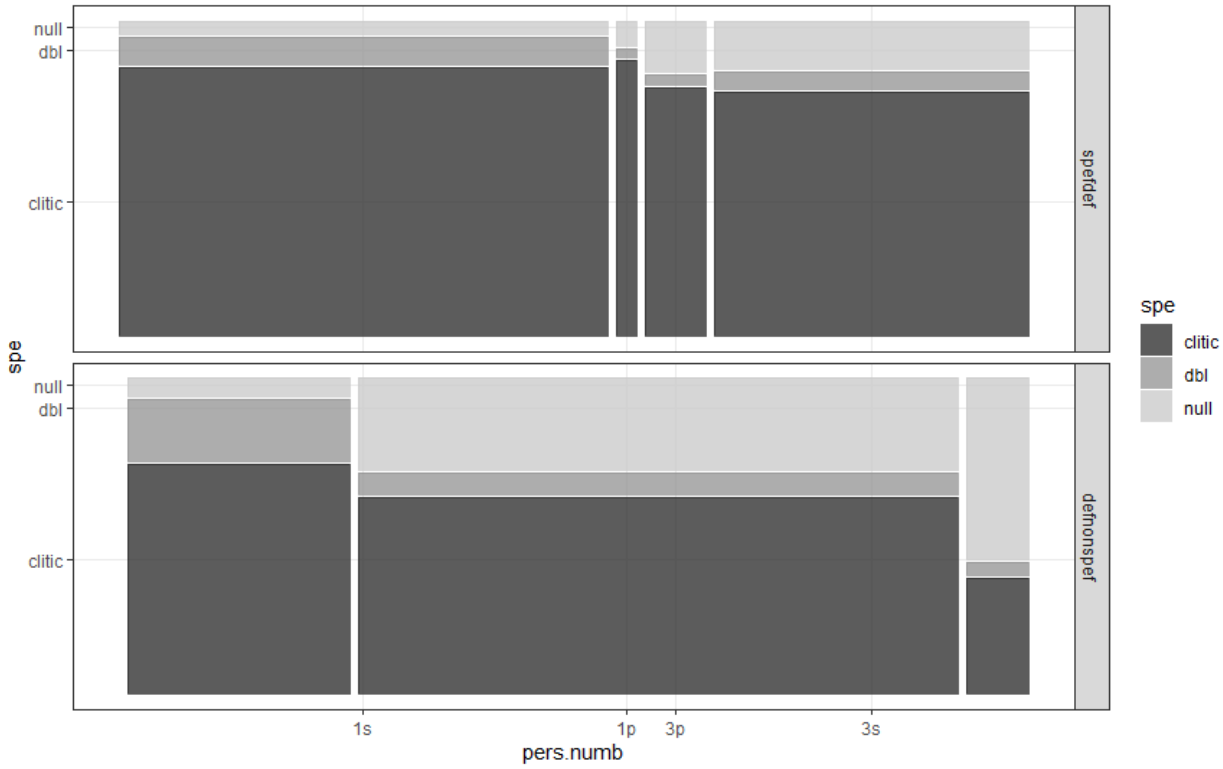
(145) *i* ***kriansa***$_i$ *ta brinka ku-el*$_j$*,* ***Ø***$_i$ *ta fazi m-****Ø***$_j$*-e di baka,* ***Ø***$_i$ *ta brinka ku-el*$_j$*,* ***Ø***$_i$ *ta toka Ø*$_j$ *dianti,* ***Ø***$_i$ *ta pega munti Ø*$_j$*,* ***Ø***$_i$ *ta inxi garrafa,* ***Ø***$_i$ *ta bira ta brinka ku-el*$_j$*.*
'and the kids play with it, [they] make it so that [it] is like a cow, [they] play with it, [they] put them up front, [they] catch a lot [of them], [they] fill the bottle, [they] play with them again.'

In the 1st person, X2SBJs were often used contrastively with nonspecific (collective, generic-like) 1PL referents. For example, speaker P19 introduced the lexical DP *Nos na rilijôn katóliku* 'us of the catholic religion' and in the next clause introduced another discourse referent *kes ki e adiventista* 'those who are Adventists'. After resuming the 3rd person plural referent *kes ki e adiventista* with 3rd person plural clitic *es*, speaker P19 reintroduced the 1st person plural referent with a X2SBJ (146).

(146) *Nada es ka ta fazi sábadu, ma **nos=nu** ta fazi.*
'They don't do anything on Saturdays, but we do [things on Saturday].'

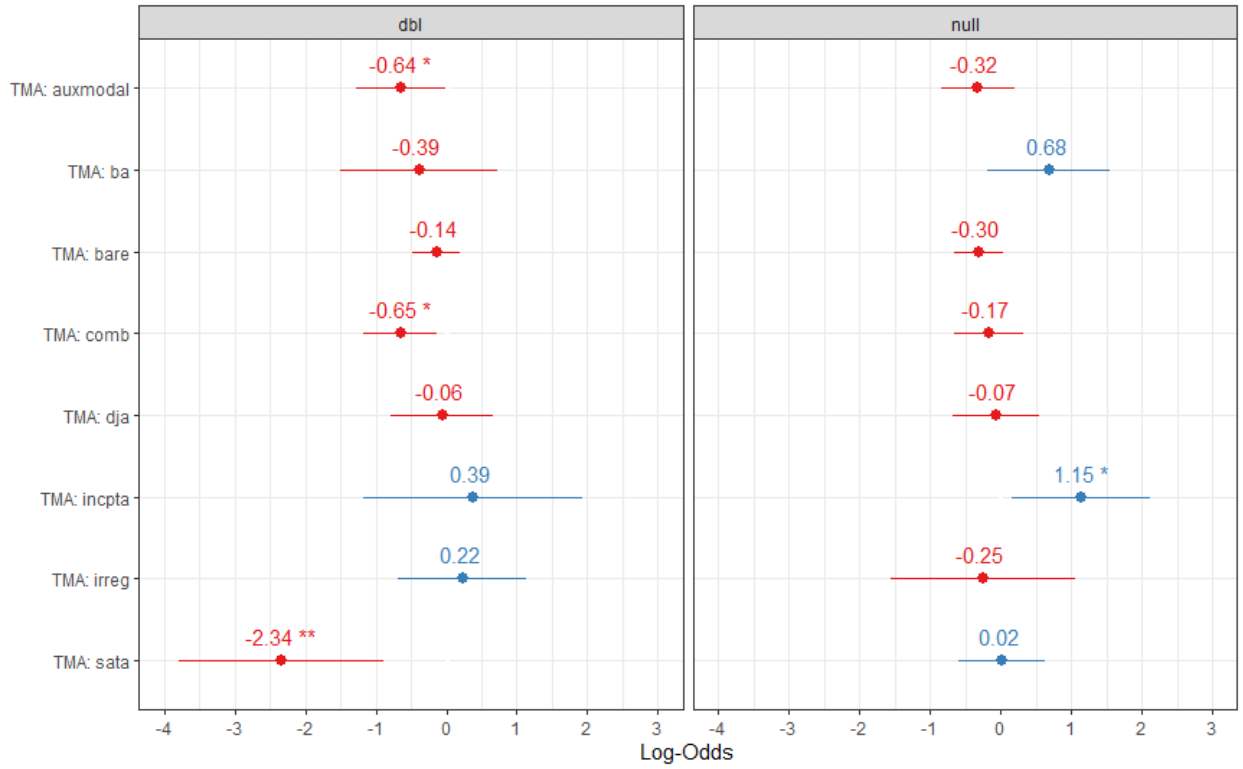
A descriptive visualization of the relationship in the mosaic plot in Figure #. In Figure #, the top panel represents SPECIFIC antecedents and the lower panel represents NONSPECIFIC antecedents, while each block represents the total count of subject realization organized by PERSON/NUMBER (x-axis) and anaphoric subject type ('clitic' (SC), double 'dbl' (X2SBJ), and 'null' (Ø)). The greater the height and width of the block, the greater portion of subject realizations. The descriptive output in Figure 33 confirms that Ø subjects were promoted most often in the 3rd person by antecedents with nonspecific reference, particularly in the picture-description narrative task. X2SBJs were more frequent in the 1st person, and especially in the 1st person plural with nonspecific (generic, collective) antecedents.

Figure 33. Mosaic plot of SPE by PERSON/NUMBER and SPECIFICITY, Analysis 2, constrained envelope.



Returning to the results from MLR 2, the factor for overt TMA marking of the verb phrase (predictor levels: BARE, AUX/MODAL, *TA*, *-BA*, *SATA*, *DJA*, INCORPORATED *TA*, IRREGULAR, COMBINATION) (Section 5.5.2.3) (Figure 34), when the target anaphor was the (non-putative, initial) subject 'slot' of the first verb in an INCORPORATED *TA* sequence, this was a favoring context for Ø subjects (LOD= 1.15; $p < 0.05$), over a SC. Progressive marker *SATA* remained a disfavoring context for X2SBJs (LOD = -2.34; $p < 0.01$), as did VPs with an AUXILIARY/MODAL (LOD = -0.64; $p < 0.05$), and those with COMBINED TMA marking, in which multiple TMA particles or TMA particles and auxiliary/modals combined (LOD = -0.65; $p < 0.05$).

Figure 34. The effects of overt TMA marking (application value = *TA*) on SPE, MLR 2, Analysis 2, constrained envelope.

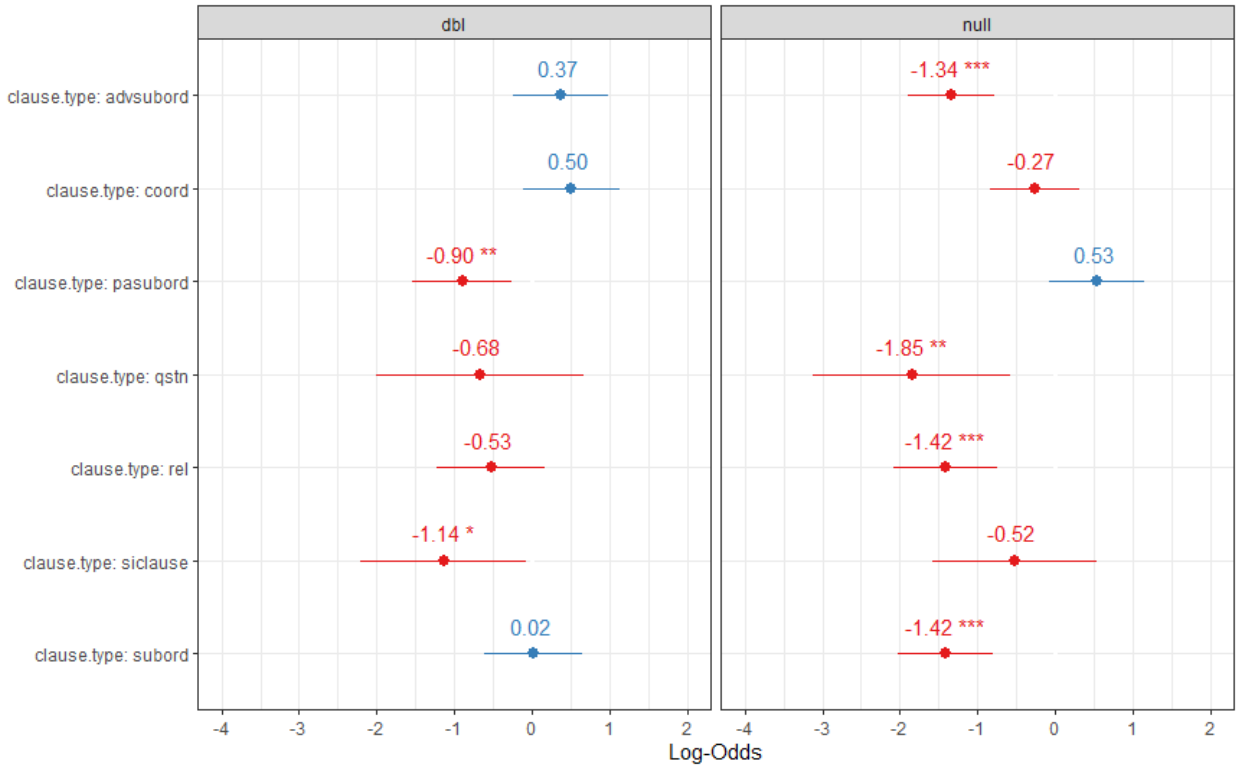


The results for TMA marking reinforce the finding that X2SBJs are disfavored in complex VPs (e.g. VPs with COMBINED TMA marking and/or modal/auxiliaries), mirroring the favoring effect that Wagner (2016) found for English Ø in complex VPs (see Section 4.5.3). X2SBJs are also strongly disfavored with progressive TMA marking, perhaps because progressive constructions might be associated with greater referential continuity. Similarly, it may be that anaphoric Ø is promoted when the target is the (non-putative) initial subject slot at the start of a serial-verb like INCORPORATED *TA* sequence; the INCORPORATED *TA* construction is also associated with a progressive or iterative interpretation and may also correlate with high referential continuity, thus promoting Ø anaphora (see Sections 4.1 and 5.4).

The removal of the COPULAR COMPLEMENT contexts from predictor CLAUSE TYPE (predictor levels = MAIN, SUBORDINATE, *PA*-SUBORDINATE, *SI*-CLAUSE, ADVERBIAL SUBORDINATE, RELATIVE, QUESTION, COORDINATE) (Section 5.5.2.1), did not substantially modify the results associated with

this factor (Figure 35). *PA*-CLAUSES were found to be a disfavoring context for X2SBJs (LOD = -0.90, p < 0.01), when compared to SCs. The strongest effect was for SUBORDINATE CLAUSES (with complementizers *ki* and *ma*) which disfavored the realization of Ø (LOD = -1.42; p < 0.001), when compared to SCs; although this context had been significant in MLR 1, it increased in significance in MLR 2 (from p < 0.01 to p < 0.001). There was a slight change from MLR 1 to MLR 2 in the magnitude of effect and overall significance, for tokens in ADVERBIAL SUBORDINATE clauses, with these targets disfavoring Ø (LOD = -1.34; p < 0.001), over SCs. RELATIVE CLAUSES became less significant and the magnitude of their effect was reduced in MLR 2, in which they continued to disfavor Ø (LOD = -1.42; p < 0.05), over SCs.

Figure 35. The effects of CLAUSE TYPE (application value = MAIN) on SPE, MLR 2, Analysis 2, constrained envelope.

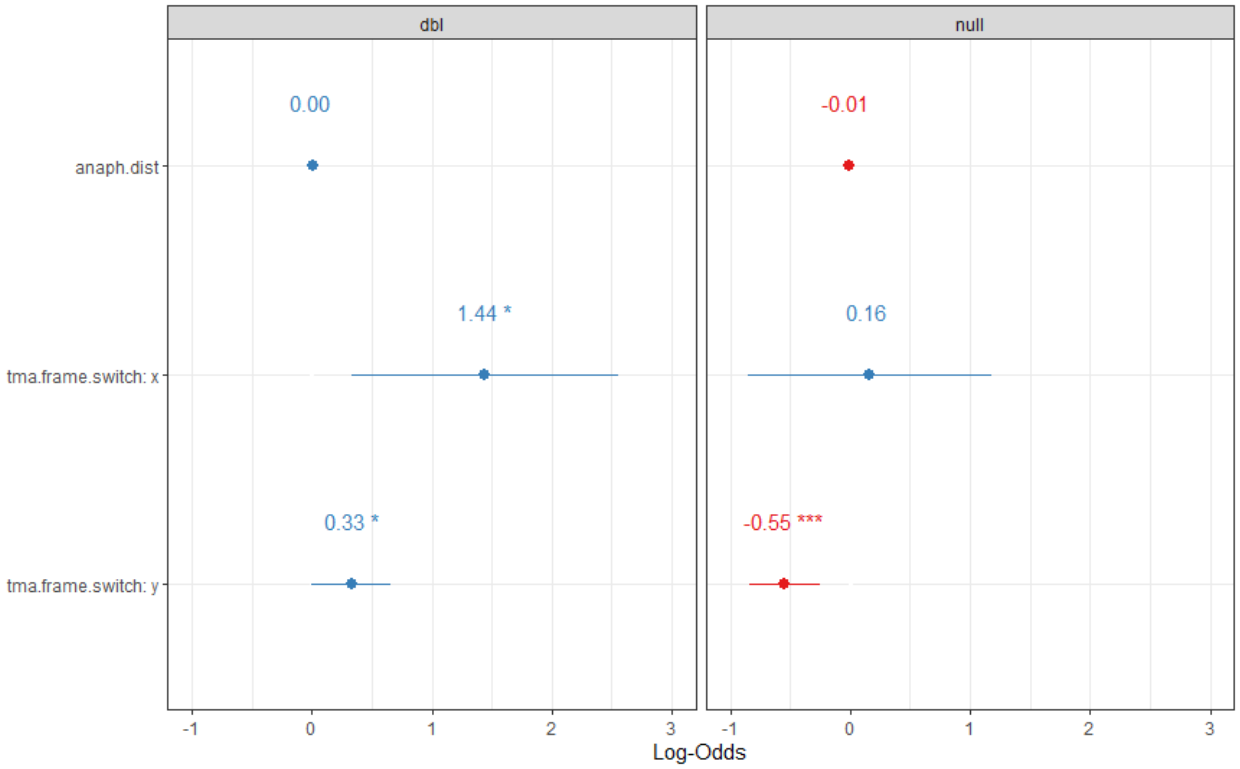


These results reinforce the observation that, in CVC, nominative anaphoric Ø is more common in main clause contexts and systematically disfavored in embedded contexts (except *pa*-

clauses), consistent with Lipski's (1999) qualitative observations for Ø subjects in several Iberian-lexifier creoles (see Section 4.5.2).

The final two language-internal factors returned significant (Figure 36) were TMA FRAME SWITCH (predictor levels = YES[=SWITCH], NO, X[=nonsubject antecedent]) (Section 5.5.2.3), and TURN, (predictor levels = RESPONSE, TURN-INITIAL, TURN-INTERNAL) (Section 5.5.2.3).

Figure 36. The effects of TMA FRAME SWITCH (application value = NO SWITCH = N) and ANAPHORIC DISTANCE (numeric continuous) on SPE, MLR 2, Analysis 2, constrained envelope.

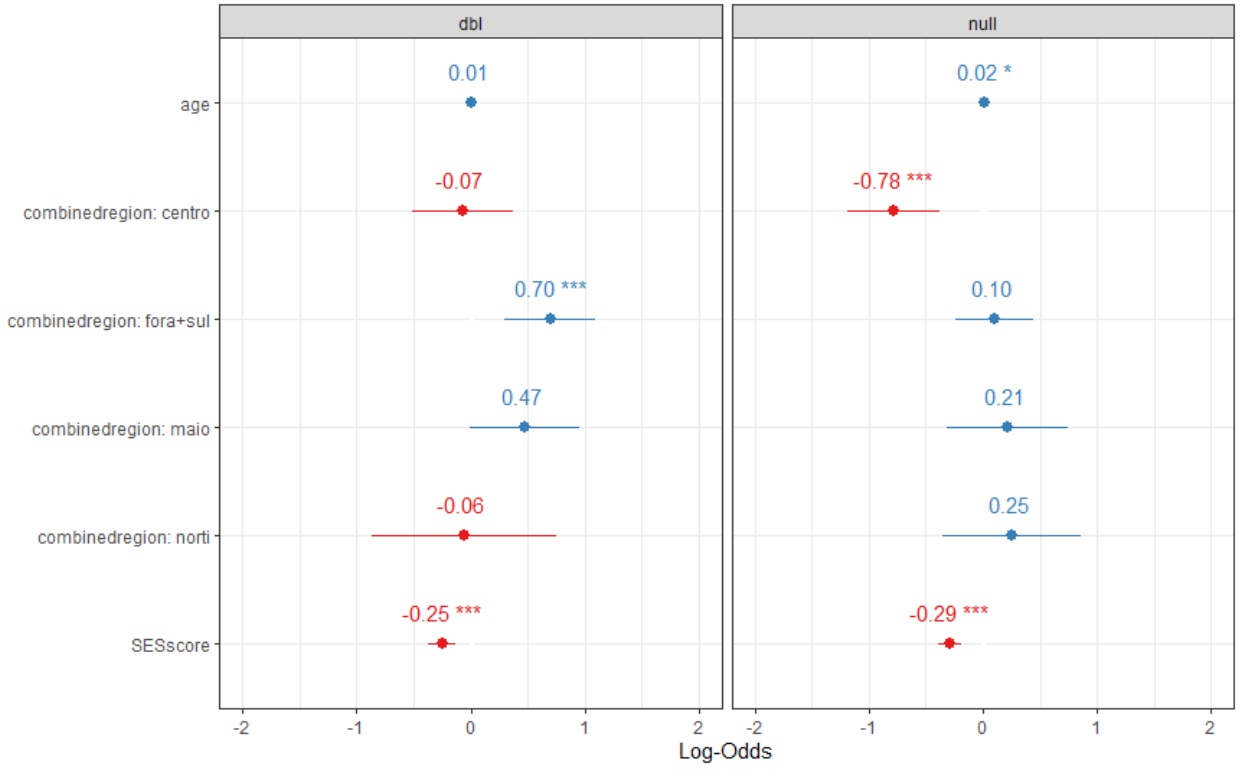


For the factor TMA FRAME SWITCH, contexts with a SWITCH (=Y) in TMA frame disfavored the realization of Ø subjects (LOD = -0.55; $p < 0.001$) and favored X2SBJs (LOD = -0.33; $p < 0.05$), when compared to SCs, respectively. Contexts in which the antecedent was not the subject in its containing clause, and thus not directly associated with the VP's TMA frame (=X), favored the realization of X2SBJs (LOD = 1.44; $p < 0.01$), when compared to SCs. This is consistent with the

notion that switches in TMA frame reduce referential continuity and thus disfavor Ø, while reinforcing the notion that X2SBJs are promoted in switch-reference contexts.

The last set of results are for the language-external individual-specific factors (Figure 37). By far the most important of these predictors was SESSCORE, which was a numeric continuous variable with values 2-8 based on the scale used in Otheguy & Zentella (2012) that assigns points to participants based on their occupation and level of education (see Section 5.5.2.4).

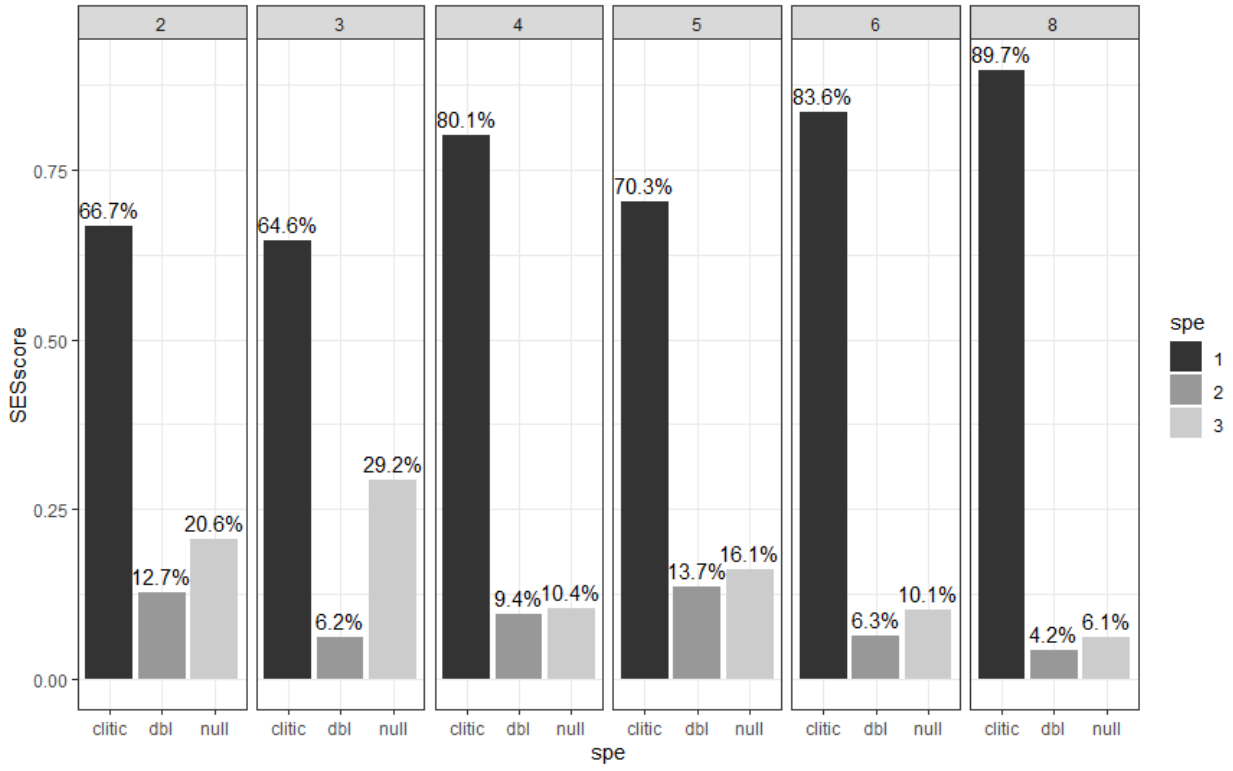
Figure 37. The effects of SESSCORE (numeric continuous, 2-8), COMBINED DIALECT REGION (application value = SUL), and AGE (numeric continuous, 18-57) on SPE, MLR 2, Analysis 2, constrained envelope.



Results for SESSCORE in Model 2 were similar to those in Model 1, a one-point increase in SESSCORE was associated with a disfavoring effect on the realization of X2SBJs (LOD = -0.25; $p < 0.001$) when compared to SCs. The same effect can be observed for Ø (LOD = -0.29; $p < 0.001$), when compared to SCs. These results reinforce the notion that speakers with lower SESSCOREs

realize a wider variety of subject forms. A descriptive visualization of the relationship between SESscore and SPE in Model 2 is shown in the bar plot in Figure 38. The issue of why speakers with higher SESSCOREs use less of a variety of subject anaphor will be discussed further in the next two subsections of this chapter when presenting the results of the binomial analyses, and in the next chapter (Discussion/Conclusions, Section 7.3).

Figure 38. Barplot of SPE by SESscore, Analysis 2, constrained envelope.



The next set of language-external individual-specific results were for COMBINED DIALECT REGION, for which the application value was SANTIAGU SUL (the region centered around the capital city Praia). *FORA* + SUL speakers, who grew-up in one of the rural dialect zones and moved to Praia in adulthood (as opposed to being born and raised and residing in SANTIAGO SUL) realized significantly more X2SBJS (LOD = 0.70; $p < 0.001$), when compared to SCS. Speakers from SANTIAGU SENTRU exhibited a disfavoring effect for Ø subjects (LOD = -0.78; $p < 0.001$) relative to SCS and when compared to speakers from SANTIAGU SUL.

Finally, with respect to AGE, a one-year increase in AGE was associated with a weak favoring effect on the realization of anaphoric Ø (LOD = 0.02, $p < 0.05$), over SCs. These results may be related to lower educational attainment in adults who received schooling before the independence of Cabo Verde in 1975 (see Sections 2.6 and 5.3.2.2.); these speakers would have lower SESSCORES and used a wider variety of subject forms.

Regression models with excessive interactions among predictor variables have high degrees of collinearity (see Section 5.7). To assess multicollinearity in Model 2, a generalized variable factor inflation diagnostic (Fox & Monette 1990) was conducted using the `vif` function in the `usdm` package (Naimi *et al.* 2017) in R (R Core Team 2019). Following the standardized 'rule of thumb' recommendations for VIF (cf. Belsley 1991:56; Kassamabara 2018), predictors with ($GVIF^{(1/2*df)}$) >5 were taken to have 'weak dependency' (some collinearity) and should have been removed from the model (Table 38). This assessment revealed SESSCORE to be the only predictor in MLR 2 approaching the threshold of five, but remained under five. The results from Analysis 2 can thus safely be assumed to be free of multicollinearity, and the results can be relied upon for drawing conclusions with confidence.

Table 38. Generalized variable factor inflation diagnostic for Analysis 2.

	GVIF	**Df**	**GVIF$^{(1/(2*Df))}$**
SURFACE FORM OF THE ANTECEDENT	5.532	7	1.130
LINKING	5.139	3	1.314
SOCIOECONOMIC STATUS SCORE	22.209	1	4.713
ANTECEDENT ACCESSIBILITY PATTERN	135.675	6	1.506
PERSON/NUMBER	13.179	3	1.537
TASK	135.675	6	1.506
TMA	7.865	8	1.138
CLAUSE. TYPE	4.150	7	1.107
SPECIFICITY	3.263	1	1.806
TMA FRAME SWITCH	28.054	2	2.301
ANAPHORIC DISTANCE	1.673	1	1.293
COMBINED REGION	37.445	4	1.573
AGE	12.670	1	3.559
VERB CLASS	3.968	3	1.258
TURN	5.114	2	1.504
CHAIN POSITION	5.538	3	1.330

Having presented the results from the fixed-effects MLR 2 in Analysis 2 representing the constrained-envelope of variation, I turn now to mixed-effects binomial logistic regression models in Analyses 3 and 4.

6.3 Analyses 3 and 4: Two mixed-effects binomial logistic regression models.

In Analysis 1, the variable importance plot of the random forest procedure indicated that a large portion of the variability in the dataset was attributable to the idiosyncratic SPE patterns of individual speakers. This indicates that statistical modeling would benefit from the inclusion of a

random factor for individual speaker. While there are several R packages capable of handling multinomial logistic regression with mixed-effects like `mlogit` (Croissant 2010) and `MCMCglmm` (Hadfield 2010), restrictions of scope and time have prevented me from analyzing the present data using these procedures[80].

To analyze the variance associated with the random factor for individual speaker (PARTICIPANT), I constructed two binomial mixed-effects logistic regressions using the command `glmer` in the `lme4` package (Bates *et al.* 2015). For the first binomial model, Model 3, in Analysis 3, I isolated instances of Ø subjects and compared these against all other nominative anaphora, collapsing SCs and X2SBJs into a single OVERT category which served as the application value. The binomial model, Model 4, in Analysis 4, collapsed Ø and SCs into a single ALL ELSE category, against which X2SBJs were compared.

For both Models 3 and 4, predictor selection and model building was carried out using stepwise logistic regression for comparing nested models' AIC scores with the `step(glm)` function in R (R Core Team 2019). Then, a 'base' model was arrived upon using the same backwards elimination procedure that was used to arrive upon the convergent multinomial models (MLRs 1 and 2).

I began with all possible parameters in an order that reflected a compromise between the parameter orders displayed in the outputs of the stepwise analysis and the variable importance plot (up to and including all the predictors retained from the stepwise procedure). Predictors that were retained first and in the majority of nested models in the stepwise procedure, and that also ranked high in variable importance in the random forest, were given precedence. Then, I began eliminating predictor terms one-by-one until model convergence was achieved. The first-to-converge model served as a 'base' against which to compare similar iterations of the model, with each iteration of

[80] See Lerche (2017) for the application of MCMCglmm to explore frequency effects in an ongoing change in German imperatives, Gray, Drummond, & Greenhill (2009) who apply the procedure for phylogenic classification and language evolution in the Pacific, and Smith & Johnson (2007) who apply the procedure to compare "Weighted vs. Probabilistic Context Free Grammars" for 'expressivity'.

the 'base' model having a (less important) predictor added or removed, and different interaction terms included, to see if they showed an improvement over the 'base' model's AIC score.

The 'base' model and its iterations were then nested once again and submitted to another stepwise comparison for AIC score using the chi-square goodness of fit test. The model with the best (lowest) AIC score was chosen for the ultimate analysis. All binomial mixed-effects models used the "BOBYQA" optimizer to facilitate model convergence[81] (Powell 2009; Bates *et al*. 2015).

The next subsection reports the results from Model 3 (Ø vs. OVERT). The next subsection after that reports the results from Model 4 (X2SBJ vs. ALL ELSE). I then conclude the chapter with a final subsection offering a summary of the results across models and concluding remarks.

Recall that this chapter is intended as a technical description of the results. The following chapter considers the aggregate of the results, discusses them, and draws final conclusions as to their meaning and relevance for the themes of this dissertation.

6.3.1 Model 3: mixed-effects binomial logistic regression, Ø vs. OVERT, constrained envelope.

The first of the two binomial models, Model 3, compared the occurrence of Ø subjects against a collapsed category representing all overt realizations (OVERT = SCS + X2SBJS). The convergent model[82] of best fit included predictors ANTECEDENT SURFACE + LINKING + PERSON/NUMBER + CLAUSE TYPE + SESSCORE + TMA FRAME SWITCH + TASK, as well as random factor (1 | PARTICIPANT).

As with the previous two Analyses, the fixed-effects estimates, p-values, and standard errors are plotted visually, while the text/numeric output of the model is provided in appendices (Appendix 4). Fixed-effects are presented first in their relative order of importance, followed by the random effects. Test for singularity and multicollinearity were conducted revealing a non-

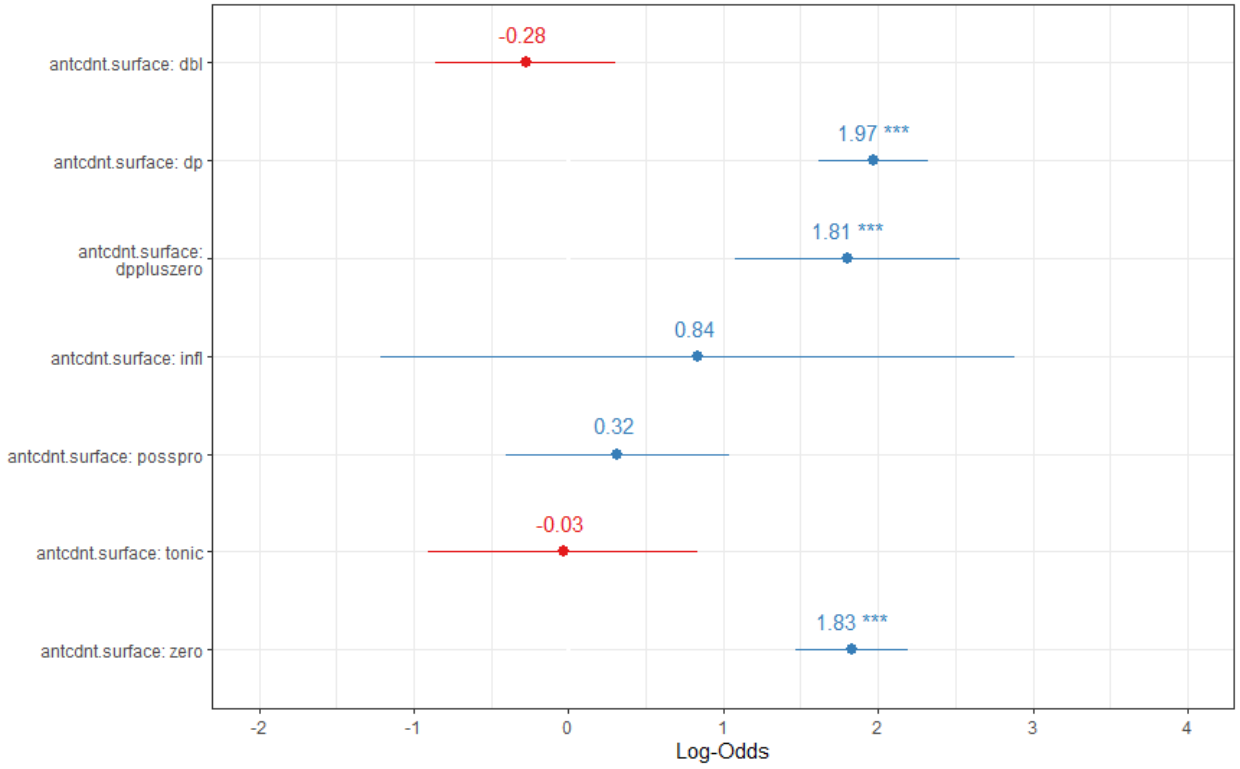
81 For use of the "BOBYQA" optimizer with mixed-effects models in linguistics see (Xia 2018; Farrington 2019; Hashimoto 2019)

82 Model AIC = 1896.3; Deviance = 1844.3 ; Residual Degrees of freedom = 3625; Random effects: Variance = 0.1434, Standard Deviation = 0.3787

singular model and no collinearity ((GVIF$^{(1/2*Df)}$) approaching 1 for all predictors) (see Appendix 5).

The results for the predictor SURFACE FORM OF THE ANTECEDENT (predictor levels = X2SBJ, LEXICAL DP, DP + INTERVENING MATERIAL + Ø, INFLECTION, POSSESSIVE PRONOUN, TONIC, Ø) (Section 5.5.2.1.), with application value CLITIC, are displayed in Figure 39.

Figure 39. The effects of SURFACE FORM OF THE ANTECEDENT (application value = CLITIC) on the realization Ø subjects over OVERT subjects, mixed-effects binomial logistic regression, Model 3, Analysis 3.

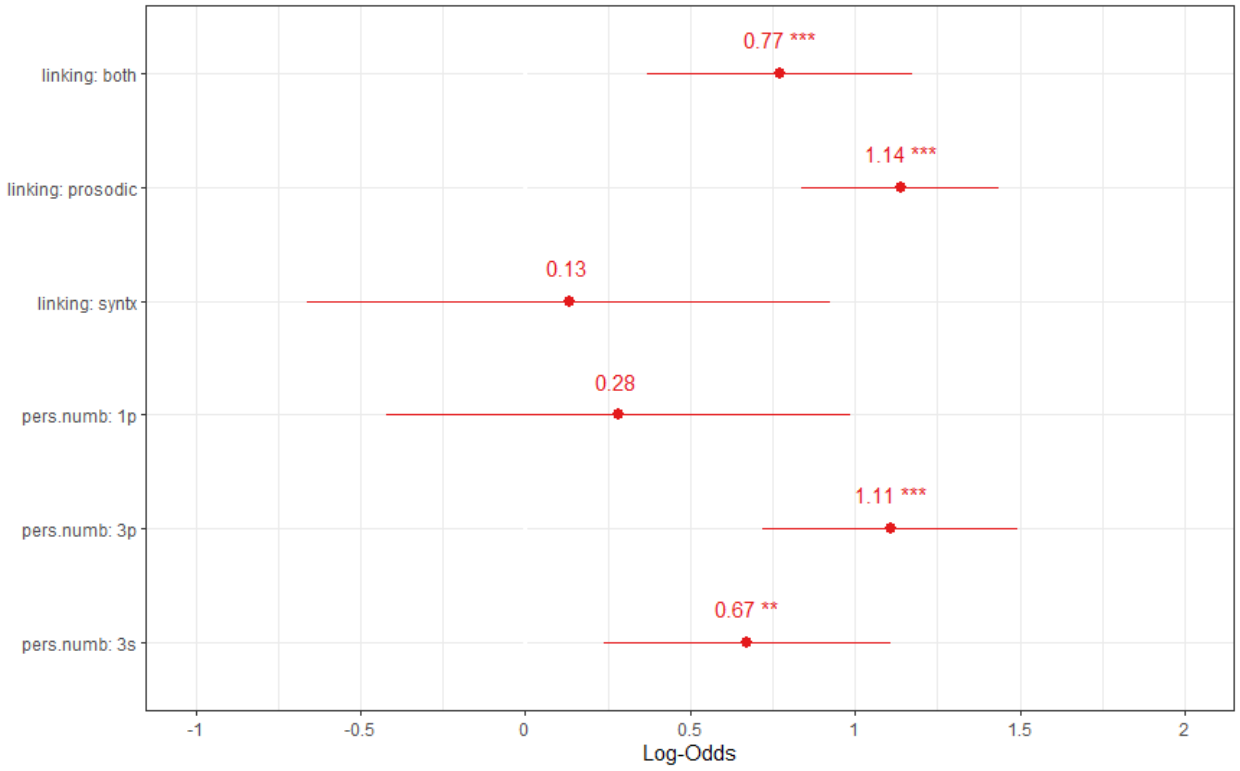


These results, including the associated significance and magnitude of the effect attributed to each predictor level, reflect closely the results for Ø subjects from the refined-envelope multinomial model (MLR 2). The Ø-to-Ø priming effect was maintained at the same significance level and with only a slightly lower LOD of the estimate (LOD = 1.83; $p < 0.001$). The favoring effect from LEXICAL DPS (LOD = 1.97, $p < 0.001$) and DP + INTERVENING MATERIAL + Ø

antecedents (LOD = 1.81, p < 0.001) on the realization of anaphoric Ø persisted when compared to OVERT forms. The LOD for the effect of LEXICAL DPs are slightly lower than in MLR 2 while the LOD for the effect of DP + INTERVENING MATERIAL + Ø antecedents are slightly higher. Overall, the results for SURFACE FORM OF THE ANTECEDENT in Model 3 support those found for MLR 2. Structural priming is the strongest effect, with Ø subjects priming subsequent Ø subjects. The same favoring effect for LEXICAL DPs and the DP + INTERVENING MATERIAL + Ø construction held as well; later in this subsection I revisit the effects for semantic-referential properties of LEXICAL DPs and similar referring expression on anaphoric Ø CVC.

The results for LINKING (PROSODIC, SYNTACTIC, BOTH, NO LINK) (Sections 4.5.3, 5.5, and 5.5.2.1), with application value NO LINK, are presented in Figure 40. There were two major changes from Model 2 in the outcomes for this predictor.

Figure 40. The effects of LINKING (application value = NO LINK) and PERSON/NUMBER (application value = 1SG) on the realization Ø subjects over OVERT pronominal subjects, mixed-effects binomial logistic regression, Model 3, Analysis 3.



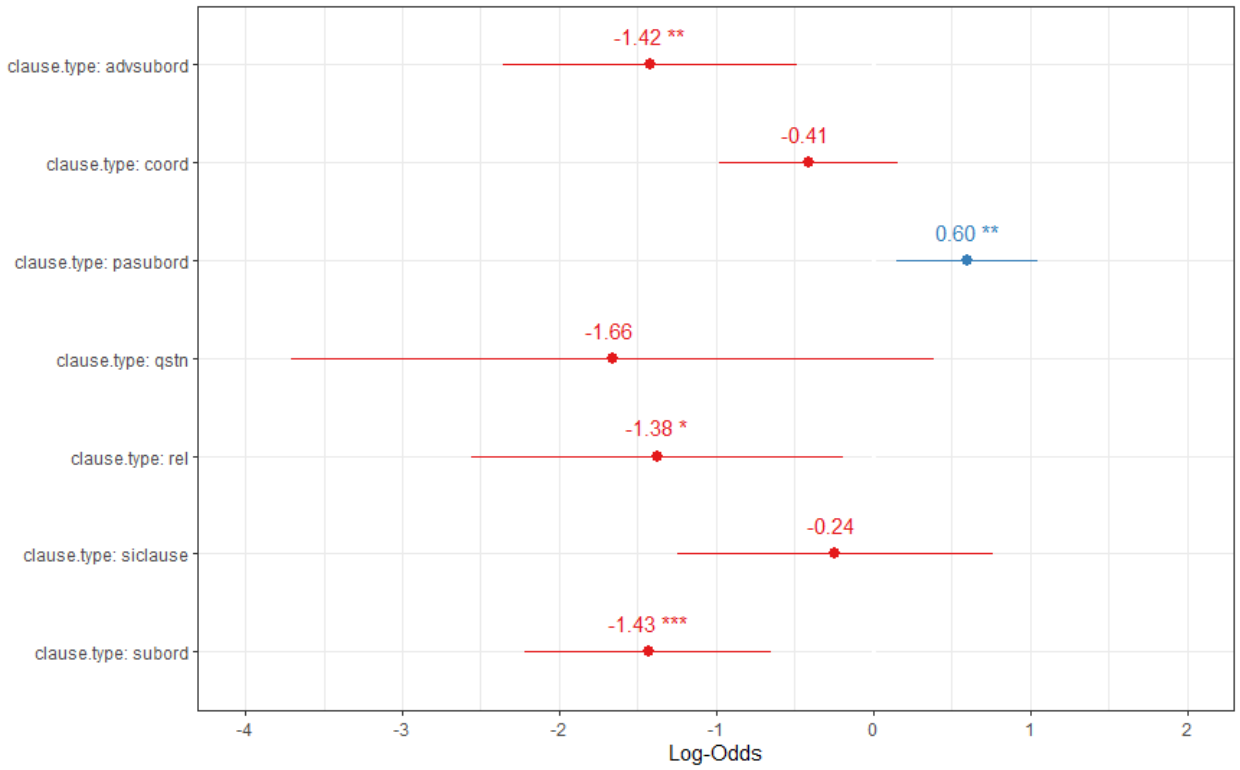
While PROSODIC LINKING remains a significant predictor of target anaphoric Ø over OVERT pronominal realizations, its overall significance has increased, and the LOD of the estimates representing the magnitude of its effect are greater than in MLR 2 (LOD = 1.14, $p < 0.001$). The level BOTH (i.e. PROSODIC and SYNTACTIC LINKING) encoded contexts where the IU containing a target subject was BOTH syntactically and prosodically linked to the immediately prior adjacent IU containing its antecedent. Unlike in MLR 2, contexts with BOTH prosodic and syntactic linking achieved significance in Model 3, exerting a strong favoring effect on the realization of Ø (LOD = 0.77, $p < 0.001$), over OVERT pronominal forms. The results for LINKING provide support for what was found in MLR 2 and add an additional significant context: Ø subjects are promoted over OVERT forms when adjacent IUs with coreferential antecedent-anaphor are prosodically linked, or simultaneously prosodically and syntactically linked. Torres Cacoullos & Travis (2019) found comparable results for English and Spanish (once the envelope of variation for SPE in each language was delimited). LINKING, and in particular PROSODIC LINKING, is a good candidate for a predictor that promotes the use of anaphoric Ø cross-linguistically. This is likely because it facilitates ANTECEDENT ACCESSIBILITY, a condition that is known to promote the use of Ø subjects over overt forms (cf. Givón 1976, Ariel 1990) (Section 4.5.3).

The predictor PERSON/NUMBER (predictor levels = 1SG, 1PL, 3SG, 3PL) (Section 5.5.2.2) had the application value 1SG (Figure 51, above). The results for PERSON/NUMBER revealed an increased significance level and greater magnitude of the effect attributable to target subjects with 3PL referents when compared to the results from MLR 2. 3PL referents exerted a strong favoring effect on the realization Ø subjects (LOD = 1.11, $p < 0.001$), when compared to OVERT subjects. 3SG referents also exerted a favoring effect on the realization Ø subjects (LOD = 0.67, $p < 0.01$), when compared to OVERT subjects. These results reinforce the observed preference for Ø subjects in the 3rd person in CVC. This reflects what has been found for Brazilian Portuguese (cf. Barbosa, Kato, & Duarte 2005; Duarte & Soares da Silva 2016).

Results for CLAUSE TYPE (predictor levels = MAIN, SUBORDINATE, *PA*-SUBORDINATE, *SI*-CLAUSE, ADVERBIAL SUBORDINATE, RELATIVE, QUESTION, COORDINATE) (Section 5.5.2.1), in MLR

2 revealed a disfavoring effect on the realization of Ø in most embedded contexts. This was reproduced in Model 3 (Figure 41).

Figure 41. The effects of CLAUSE TYPE (application value = MAIN) on the realization Ø subjects over OVERT subjects, mixed-effects binomial logistic regression, Model 3, Analysis 3.



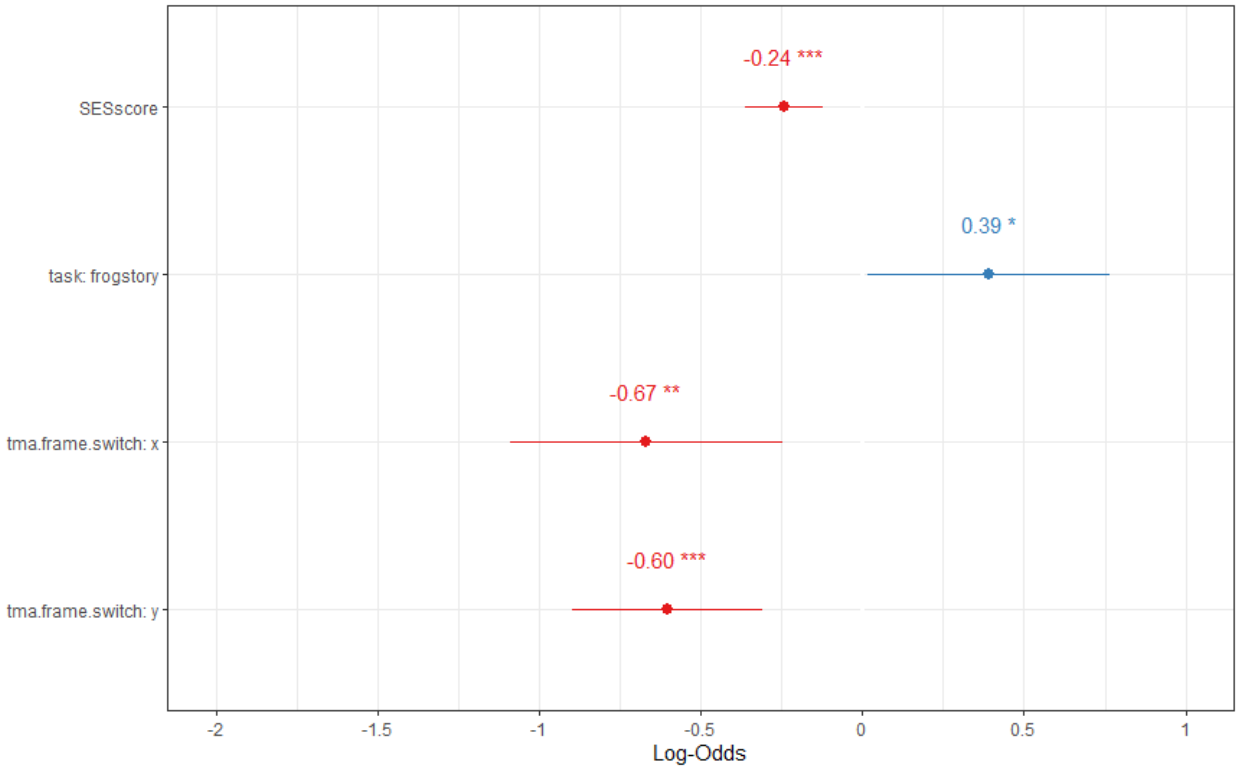
There was a strong disfavoring effect for Ø in SUBORDINATE clauses (after complementizers *ki* and *ma*) (LOD = -1.43; $p < 0.001$) and ADVERBIAL SUBORDINATE clauses (LOD = -1.42; $p < 0.001$), and a weaker effect was found for Ø in RELATIVE clauses (LOD = -1.38; $p < 0.05$), when compared OVERT targets, respectively. A favoring effect for anaphoric Ø emerged in *PA*-CLAUSES (LOD = 0.60, $p < 0.01$), relative to OVERT pronominals. These results support the general notion that, in CVC, just as Lipski (1999) observed for several other Iberian-origin creole languages, Ø subjects tend to occur in main clause contexts. While Lipski called into question the genuine argumental status of the Ø subjects he was observing (except in Chabacano), the Ø

subjects in question here for CVC are genuinely anaphoric, as they are all tied to some discernable antecedent (referentially fully specified OR deficient).

The results for *pa*-clauses reflect what Kouwenberg (1990) found for Papiamentu; anaphoric Ø can alternate with a quasi-argumental or 'generic' Ø (pro_{ARB}) in these contexts, and this variability likely reflects the 'scalar' reduction in 'finiteness' (since finiteness is not a binary category in CVC and many other languages, cf. Givón 2001[1984]; Estrada Fernández 2016) in embedded *pa*-clauses (see Section 5.4). Given that Jacobs (2010; 2012) and Jacobs & Quint (2016) have provided strong evidence that CVC and Papiamentu belong to the same Upper Guinea Creole Portuguese group of languages (see Section 2.3), it is no surprise that embedded *pa*-clauses should appear to pattern similarly in these varieties. A close comparison of Ø subjects in *pa*-clauses in CVC and Papiamentu is promising topic for future analysis.

The final three fixed-effects in Model 3 were for TMA FRAME SWITCH (predictor levels = YES, NO, X[=non-subject antecedent]) (Section 5.5.2.4) and the two language-external factors: TASK (predictor levels = INTERVIEW, PICTURE DESCRIPTION NARRATIVE), and SESSCORE (numeric, continuous, 2-8) (Figure 42). The former is reported first.

Figure 42. The effects of TMA FRAME SWITCH (application value = NO SWITCH), TASK (application value = INTERVIEW), and SESSCORE (numeric continuous 2-8) on the realization Ø subjects over OVERT subjects, mixed-effects binomial logistic regression, Model 3, Analysis 3.



There was a significant disfavoring effect on Ø in TMA SWITCH (=Y) contexts (LOD = -0.60, $p < 0.001$) and in contexts in which the antecedent had some role in the clause other than that of the subject (=X) (LOD = -0.67, $p < 0.01$), when Ø was compared to OVERT pronominal subjects, respectively. The results for TMA FRAME SWITCH confirm those from MLR 2; switches in TMA frame disfavor the realization of Ø subjects relative to overt pronominals. Bayley & Pease-Álvarez (1997) and Geeslin & Gudmestad (2011) have found parallel results for the effect of TMA FRAME SWITCH on SPE in Spanish (Section 4.5.3). This effect is likely due to a decrease in antecedent accessibility incurred when switching TMA frames from one clause to the next.

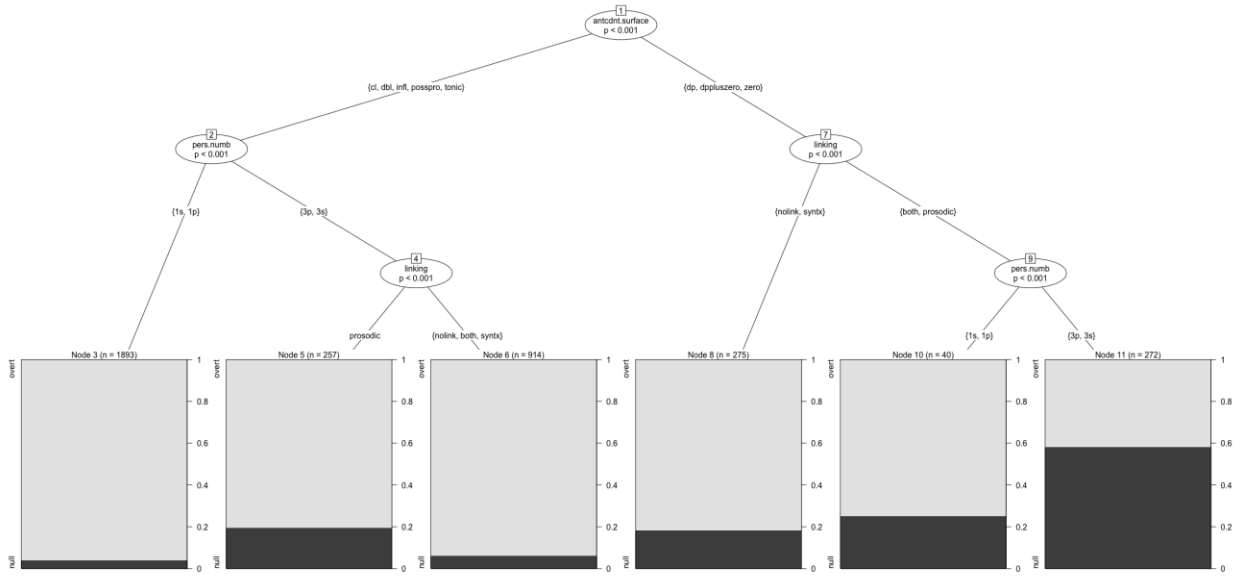
For the predictor TASK, as in MLR 2, the PICTURE DESCRIPTION NARRATIVE ('the Frog story') was a favoring context for Ø (LOD = 0.39, $p < 0.05$), over OVERT pronominal subjects.

This may be related to the subject continuity associated with the narrative in which only three characters are involved, requiring less back-and-forth switching and reintroduction of old discourse referents, as well as the higher rate of 3rd person referents when compared to the interview task. The possibility that this effect has more to do with narrative speech style (than some property of morphosyntactic composition) should not, however, be discarded.

Finally, as in MLR 2, an increase of a magnitude of one in a participant's SESSCORE (numeric continuous, 2-8) (Section 5.5.2.4) was associated with a disfavoring effect on the realization for Ø (LOD = -0.24, $p < 0.001$), when compared to OVERT pronominal. This result lends support to the observed trend in the MLR models whereby higher SESSCOREs were associated with less of a variety of subject forms and higher rates of SCs.

To more closely inspect interactions among the fixed effects, several conditional inference trees were grown and predicted probabilities were plotted. For the series of predicted probabilities plots, the effects of two or three factors on SPE (all others held constant) are examined. The first conditional inference tree (Figure 43) examines the grouping of the predictors SURFACE FORM OF THE ANTECEDENT, LINKING, and PERSON/NUMBER for their effects in determining the distribution of Ø subject realizations.

Figure 43. Conditional inference tree for the outcome Ø subject over OVERT subject for the predictor terms SURFACE FORM OF THE ANTECEDENT, LINKING, and PERSON/NUMBER, Analysis 3.



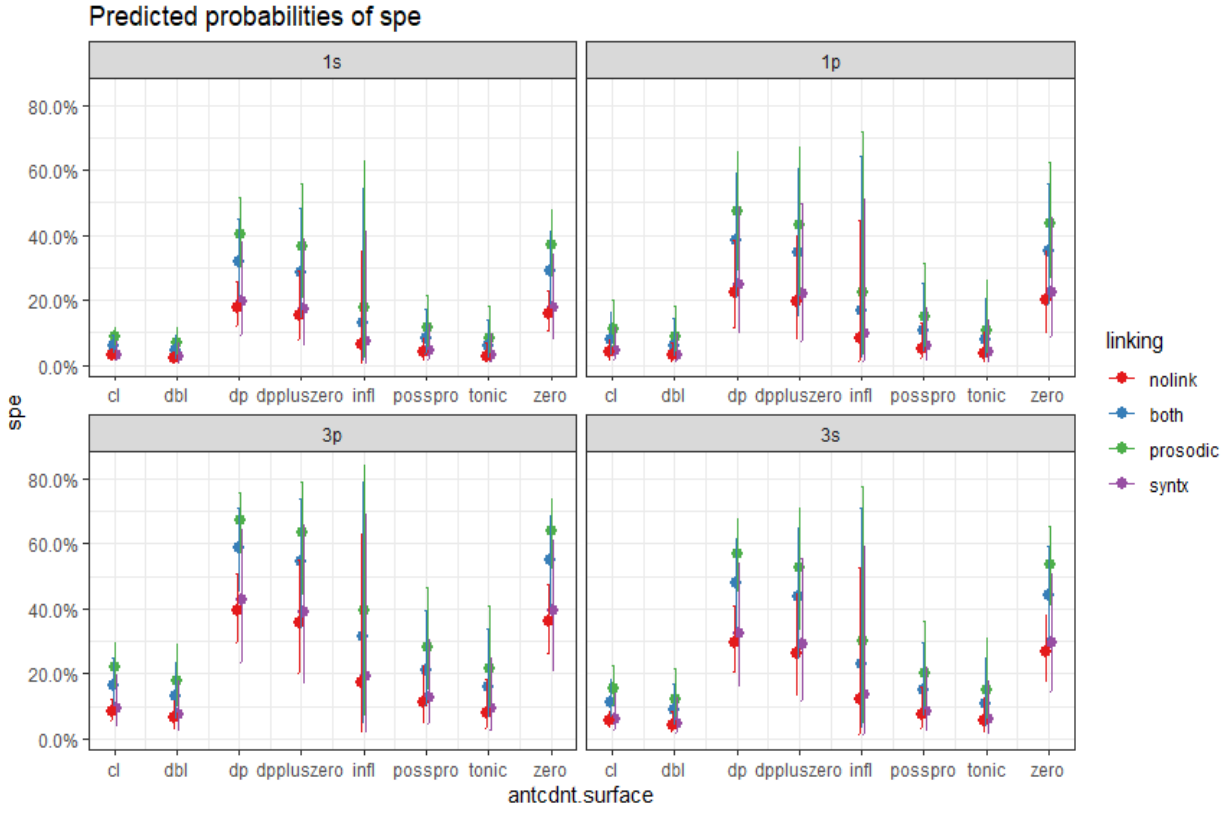
Once again, LEXICAL DP, DP + INTERVENING MATERIAL + Ø, and Ø antecedents patterned apart from all other antecedent types ($p < 0.001$). The favoring effect from these three antecedents have been found in the regression and conditional inference tree outputs from Analyses 2 and 3. Figure # shows that when these antecedents were in IUs that were prosodically linked, or BOTH prosodically and syntactically linked to the anaphor's IU, PERSON/NUMBER became relevant ($p < 0.001$). In these contexts, targets with 3rd person reference resulted in more Ø subjects and this was the most favorable context for anaphoric Ø. The second most favorable context for Ø was the equivalent context but with 1st person referents. When these three antecedents were unlinked to their coreferential target anaphora, or when they were only syntactically linked to their containing clauses, PERSON/NUMBER ceased to be relevant. According to these results, LINKING can be viewed as 'unlocking' the person-number effect on SPE.

For all other antecedent types, subjects with 3rd person referents patterned apart from those with 1st person referents ($p < 0.001$). For all other antecedent bearing 3rd person reference, PROSODIC linking was associated with higher rates of Ø ($p < 0.001$), as opposed contexts with

SYNTACTIC LINKING, BOTH prosodic and syntactic linking, or unlinked contexts. For 'all other' antecedent types bearing 1st person reference, LINKING was not relevant, and overall this was the least favorable context for Ø.

The predicted probabilities of realizing Ø as an effect of SURFACE FORM OF THE ANTECEDENT (plotted on the x-axes), LINKING (dots and error bars), and PERSON/NUMBER (panels) are plotted in Figure 44. It can be seen that the most likely context for the realization of Ø over an OVERT pronominal subject was with LEXICAL DPs, the DP + INTERVENING MATERIAL + Ø construction, and Ø antecedents, across prosodically LINKED IUs, with 3PL referents, and only slightly less so with 3SG referents.

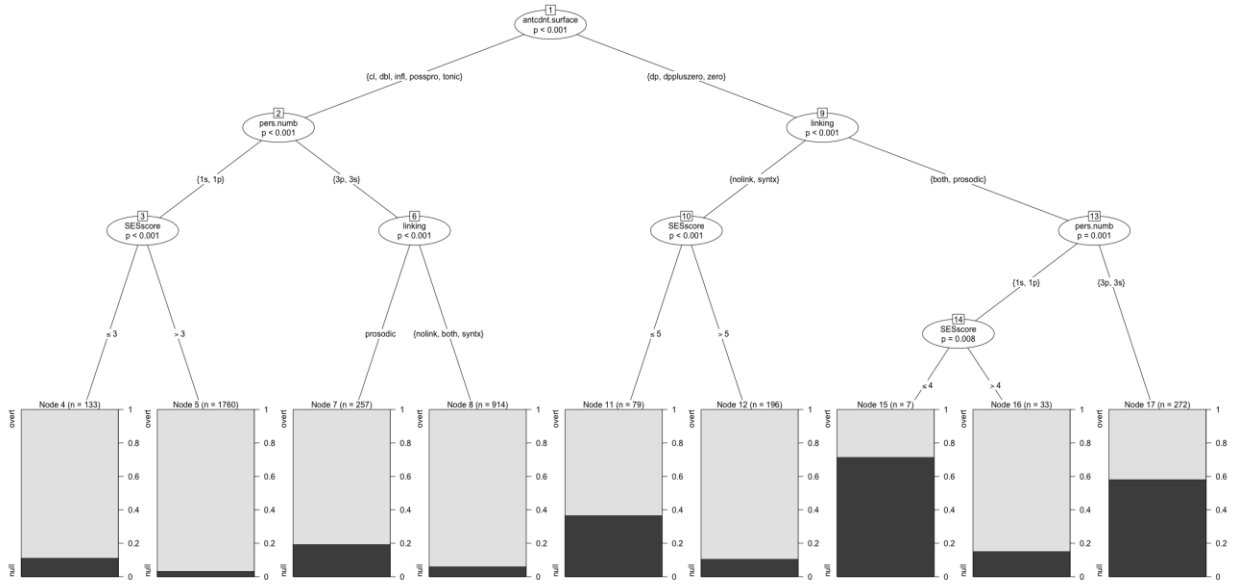
Figure 44. Predicted probabilities of realizing Ø subjects over OVERT subjects as an effect of SURFACE FORM OF THE ANTECEDENT, LINKING, and PERSON/NUMBER, mixed-effects binomial logistic regression, Analysis 3.



These results support the notion that LINKING enhances referential continuity, and therefore increases antecedent accessibility; LINKING allows the Ø-to-Ø priming effect to obtain, and also activates the influence of LEXICAL DP antecedents with referentially deficient semantic values on the realization of Ø subjects.

A conditional inference was grown to examine the interaction between SESSCORE and language internal factors such SURFACE FORM OF THE ANTECEDENT and PERSON/NUMBER. In the conditional inference tree (Figure 45), it can be seen that any significant effects for SESSCORE emerge in 1st person contexts, in NO LINK contexts, or in contexts with merely SYNTACTIC LINKING, but never in 3rd person contexts or contexts with PROSODIC LINKING or BOTH prosodic and syntactic linking.

Figure 45. Conditional inference tree for SPE (Ø vs. OVERT) with predictors SURFACE FORM OF THE ANTECEDENT, LINKING, PERSON/NUMBER and SESSCORE, Analysis 3.



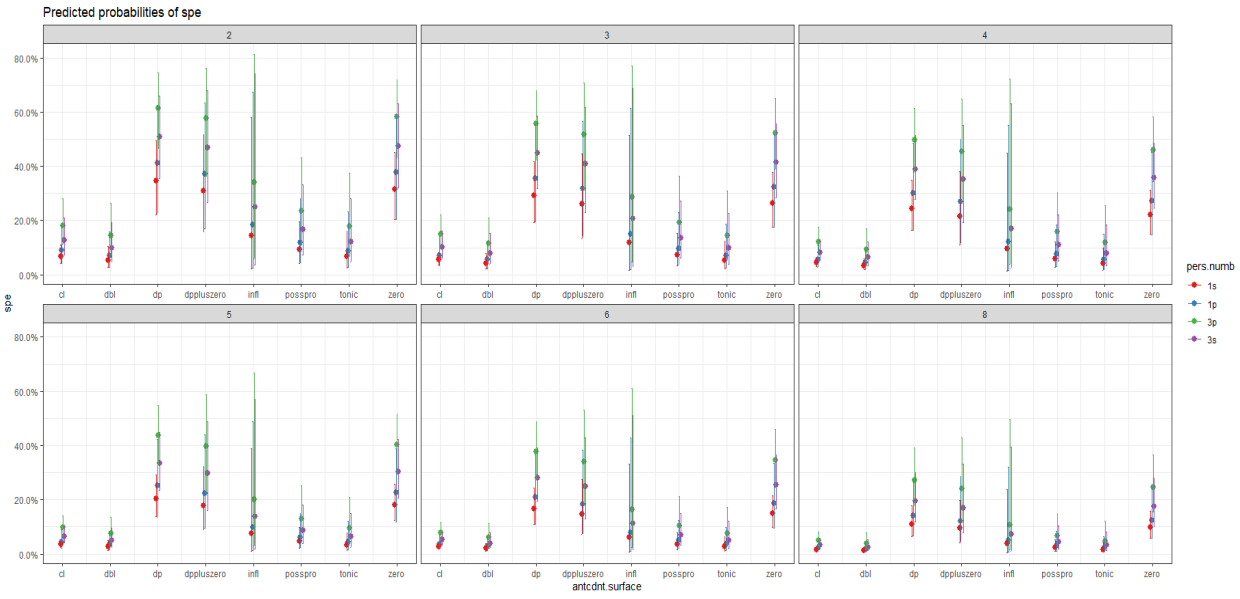
For target anaphora that bore 1st person reference and LEXICAL DP, DP + INTERVENING MATERIAL + Ø, or Ø antecedents across linked IUs, an SESSCORE of four was significant ($p < 0.01$); speakers who received an SESSCORE ≤ 4 allowed for ample anaphoric Ø with 1st person referents. When these three antecedent types remained unlinked, or were merely syntactically linked to the

anaphora's IU, an SESSCORE of five was significant ($p < 0.001$); subjects who received an SESSCORE ≤5 realized somewhat more anaphoric Ø. For 'all other' antecedent types, SESSCORE was only significant when these bore 1st person reference ($p < 0.001$), with participants who received an SESSCORE ≤3 realizing slightly more anaphoric Ø ($p < 0.001$).

These results indicate that patterns of SPE produced by speakers with lower SESSCORES are less restricted by constraints like PERSON/NUMBER and LINKING. In other words, the distribution of anaphoric Ø is sensitive to the same set of underlying constraints for all speakers, but the restrictive effect of these constraints is stronger for speakers with higher SESSCOREs. Meanwhile, speakers with lower SESSCOREs are more permissive of Ø subjects in UNLINKED and 1st person contexts.

Similar trends can be observed in the predicted probabilities of realizing Ø over OVERT pronominal subjects, as an effect of SURFACE FORM OF THE ANTECEDENT (plotted on the x-axes), PERSON/NUMBER (plotted as points and standard error lines), and SESSCORE (in facetted panels), (Figure 46).

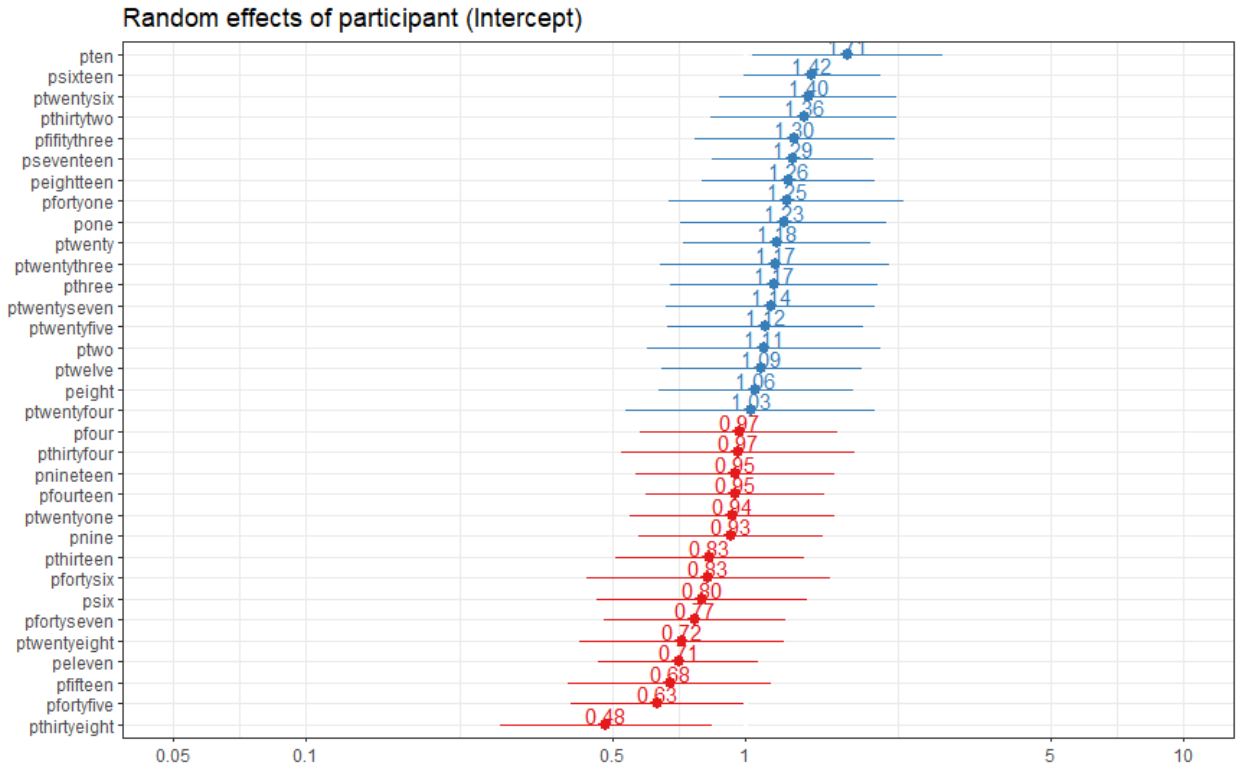
Figure 46. Predicted probabilities of realizing Ø subjects over OVERT subjects as an effect of SURFACE FORM OF THE ANTECEDENT by PERSON/NUMBER and SESSCORE, mixed-effects binomial logistic regression, Model 3, Analysis 3.



From the predicted probabilities, it can be observed that the most likely context for anaphoric Ø is when targets bore 3rd person reference, the antecedent was a LEXICAL DP, DP + INTERVENING MATERIAL + Ø construction, or Ø, and the speaker had a lower SESSCORE.

The caterpillar plot in Figure 47 displays the conditional modes (BLUPs) associated with each PARTICIPANT and their corresponding confidence intervals. The errors bars (for confidence intervals) that do not overlap with the intercept line (1) differ significantly from the population in patterns of SPE. The descriptive results for SPE realized by these participants are briefly examined.

Figure 47. Plot of the conditional modes (BLUPs) for random factor PARTICIPANT, mixed-effects binomial logistic regression, Model 3, Analysis 3.



Speaker P10, at the upper bounds of the plot, differed from the population average in realizing higher rates of Ø subjects. Speaker P10 was 37 at the time of interview, was born-and-raised in Praia, and received an SESSCORE of 3. P10 produced high rates of Ø relative to the total number of tokens isolated from their speech: 19 of 46 tokens (=29.2%) were anaphoric Ø. The

results for the random factor PARTICIPANT indicate that pattern of Ø subjects do not appear to be especially sensitive to idiosyncratic patterns of SPE associated with individual participants.

In this section, it was shown that many of the results for Ø subjects in MLR 2 persisted in Model 3 for Ø subjects. Priming remained the most important and strongest effect, as did the effect for the relationship between lexical DPs with nonspecific reference and anaphoric Ø. Both these effects are enable when adjacent IU containing antecedent and anaphor are prosodically linked, or prosodically and syntactically linked. Linking also promotes Ø subjects when the referent is 3rd person, and speakers with higher SESSCOREs are stricter in their application of constraints like LINKING on patterns of SPE. The generalized disfavoring effect on Ø in embedded contexts is confirmed, with the exception of *pa*-clauses, which like in Papiamentu allow for variation between anaphoric Ø, generic/quasi-argumental Ø (*pro*$_{ARB}$), and overt pronominals. Switches in TMA frame from clause to clause disfavor Ø subjects, probably because of the reduction incurred in referential continuity and thus antecedent accessibility. The factor TASK revealed that Ø was favored in the PICTURE-DESCRIPTION NARRATIVE, but it is unclear if this is because there are more 3rd person referents (and less overall discourse referents) throughout this task when compared with the INTERVIEW, or if this effect is due stylistic difference between a story-telling and conversational mode in CVC. I turn now to the next binomial model, Model 4, which compared X2SBJ realizations against a merged SC + Ø (ALL ELSE) category.

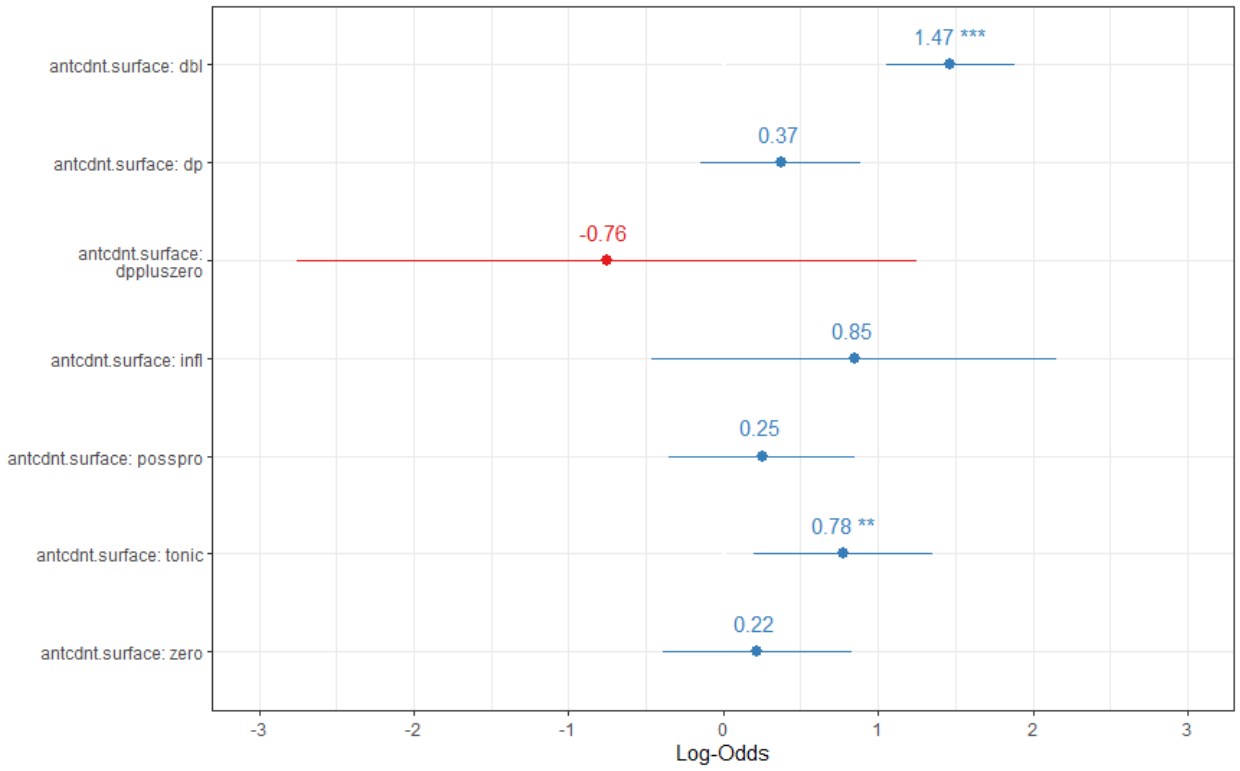
6.3.2 Model 4: mixed-effects binomial logistic regression, constrained envelope, X2SBJ vs. ALL ELSE, constrained envelope.

The fourth and final model, Model 4, was also a mixed-effects binomial logistic regression, but Model 4 compared X2SBJ realizations against ALL ELSE (SCs and Ø). The model building and selection procedure was identical to that of Model 3 (Section 6.3.1). The fixed-effects for Model 4 included SURFACE FORM OF THE ANTECEDENT, ANTECEDENT ACCESS PATTERN, TURN, SESSCORE,

and CLAUSE CHAIN POSITION, and the random intercept was PARTICIPANT[83]. Tests for singularity and multicollinearity (Appendix 7) were conducted revealing a non-singular model with no collinearity associated with the model parameters.

The first set of results (Figure 48) were for SURFACE FORM OF THE ANTECEDENT (predictor levels = X2SBJ, LEXICAL DP, DP + INTERVENING MATERIAL + Ø, INFLECTION, POSSESSIVE PRONOUN, TONIC, Ø) (Section 5.5.2.1). The priming effects for X2SBJs found in MLR 2 were maintained in Model 4; a X2SBJ antecedent exerted a favoring on a X2SBJ target (LOD = 1.47, $p < 0.001$), as did TONIC pronominals (LOD = 0.78, $p < 0.01$), when these were compared to ALL ELSE realizations, respectively.

Figure 48. The effects of SURFACE FORM OF THE ANTECEDENT (application value = CLITIC) on the realization X2SBJs over ALL ELSE (scs and Ø subjects), mixed-effects binomial logistic regression, Model 4, Analysis 4.



[83] AIC = 1679.9; Deviance = 1637.9; Random effects: Variance = 0.259; Random effects standard deviation = 0.509; Mean residuals = -0.159

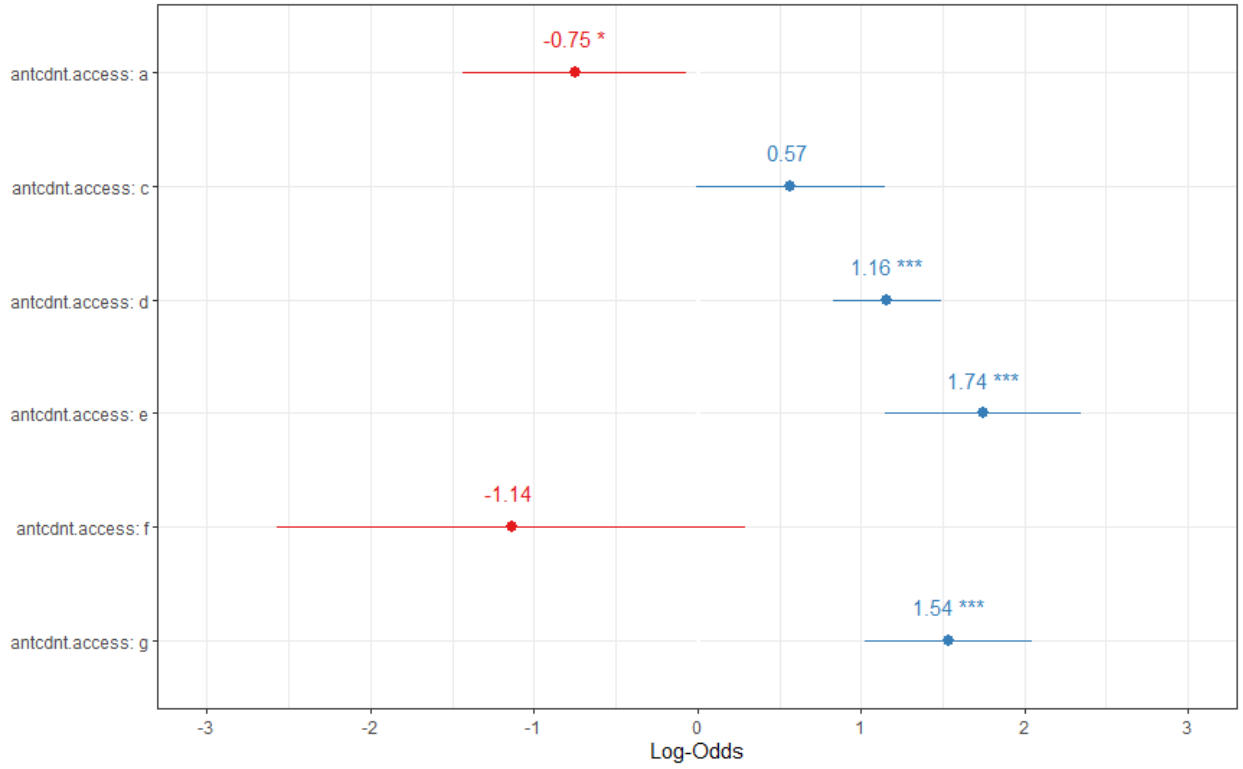
These results support the effects found for priming, which arise consistently across analyses and modes of analysis for SPE in CVC. This adds to already robust evidence that priming is a cross-linguistically relevant constraint on anaphora resolution (Cameron 1995; Flores-Ferrán 2002; Cameron & Flores-Ferrán 2004; Travis 2007; Claes 2017; Wagner 2016; Bouchard 2018; Carvalho & Child 2011; Carvalho & Bessett 2015; Abreu 2012; Torres Cacoullos & Travis 2015; *inter alia*) (Section 4.5.3). This is likely because priming is related to domain-general cognitive processes that intervene to modulate anaphora resolution (cf. Pickering & Ferreira 2008; Claes 2017).

The results for ANTECEDENT ACCESSIBILITY PATTERN (PATTERNS A, B, C, D, E, F, G; see Sections 4.5.1 and 5.5.2.1), demonstrate comparably strong effects to those found for SURFACE FORM OF THE ANTECEDENT. All switch-reference contexts in which at least one clause with a non-coreferential subject interceded between a target anaphor and its antecedent achieved significance (Figure 49). The only switch-reference context that did not achieve significance was PATTERN C, which was when the antecedent was in the immediately prior clause but had some role other than that of the subject. Therefore, it can be said that X2SBJS are favored when the antecedent is not in an adjacent IU/clause, or in other words, X2SBJS are used to reintroduce old (and anaphorically distant) discourse referents.

The strongest favoring effect was for PATTERN E (LOD = 1.74, $p < 0.001$), which was when the antecedent was separated from the target anaphor's clause by at least one clause containing a non-coreferential subject, and the antecedent played a role in its containing clause other than that of the subject (like PATTERN C but in a non-adjacent clause). The next strongest favoring effect was for PATTERN G (LOD = 1.54, $p < 0.001$), which was when the antecedent occurred in a prior, separate 'discourse chunk' from that of the target (the referent is an old discourse topic being reintroduced from a previous 'discourse chunk'). The weakest favoring context was PATTERN D (LOD = 1.16, $p < 0.001$); PATTERN D was when the antecedent was the subject of a non-adjacent IU/clause and at least one IU/clause with a non-coreferential subject intervened (like PATTERN B but in a non-adjacent clause). PATTERN A contexts disfavored the realization of X2SBJS (LOD = -

0.75, $p < 0.05$), when compared to ALL ELSE. PATTERN A referred to when the target anaphor was c-commanded by its antecedent.

Figure 49. The effects of ANTECEDENT ACCESSIBILITY PATTERN (application value = PATTERN B) on the realization X2SBJs over ALL ELSE (scs and Ø subjects), mixed-effects binomial logistic regression, Model 4, Analysis 4.

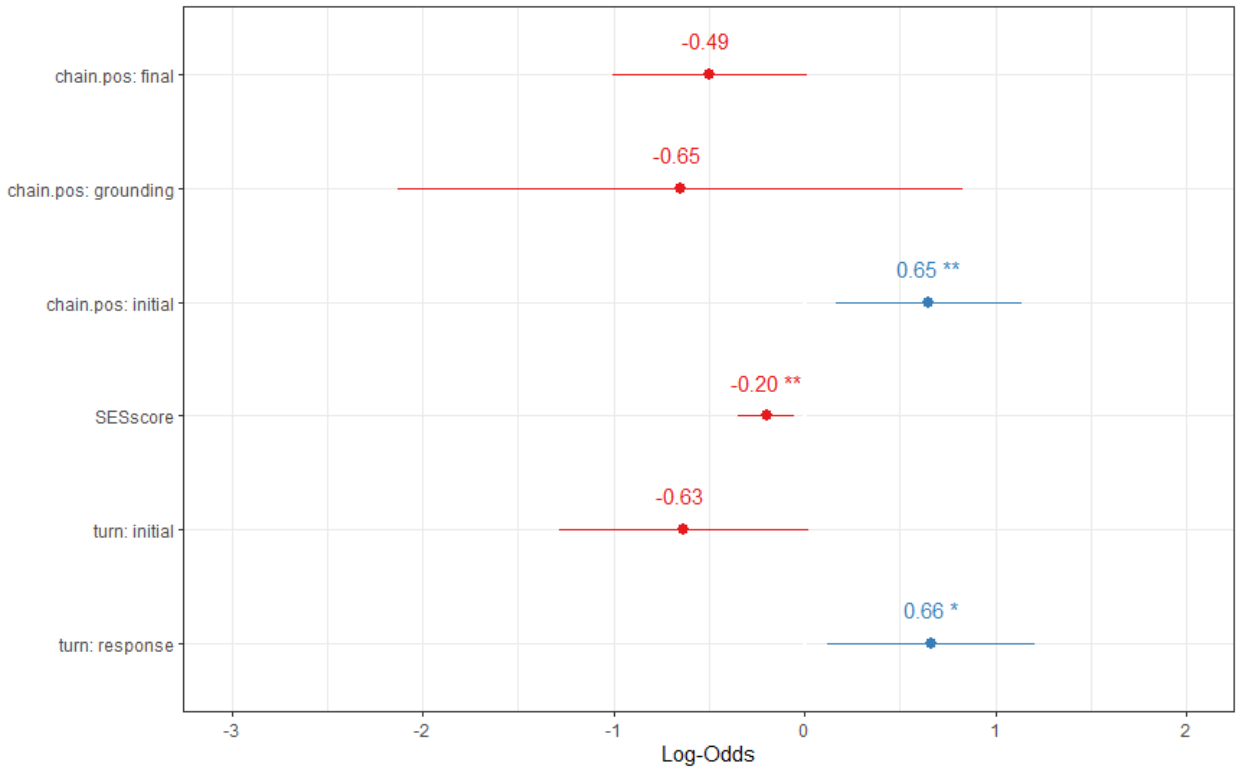


The results show that X2SBJs are most likely to occur when the antecedent is in a nonadjacent clause/IU, and therefore the discursive function of X2SBJs is to reintroduce old (anaphorically distant) discourse referents. This also lends further support to the notion that, since X2SBJs are the primary way to deploy a TONIC pronoun in CVC, their discursive function (and priming values) are essentially equivalent to those of singleton TONIC pronouns.

The next set of results (Figure 50) were for the language-internal factor CLAUSE CHAIN POSITION CHAIN POSITION (predictor levels = GROUNDING, INITIAL, MEDIAL, FINAL) (Sections 4.5.3 and 5.5.2.1), the language-external task related factor TURN (predictor levels = RESPONSE, TURN-

INITIAL, TURN-INTERNAL) (Section 5.5.2.3), and the language-external individual-specific factor SESSCORE (numeric continuous 2-8) (Section 5.5.2.4).

Figure 50. The effects of turn (application value = TURN-INTERNAL), SESscore (numeric continuous 2-8), and clause chain position (application value = CLAUSE CHAIN MEDIAL), on the realization X2SBJs over ALL ELSE (SCs and Ø subjects), mixed-effects binomial logistic regression, Model 4, Analysis 4.



With respect to TURN, when compared with TURN INTERNAL contexts, targets contained in RESPONSEs exerted a favoring effect on the realization of X2SBJs (LOD = 0.66, $p < 0.05$), as opposed to ALL ELSE (SCs and X2SBJs). This means that X2SBJs are used in response to questions, either contrastively or to promote the topic-status of a discourse referent.

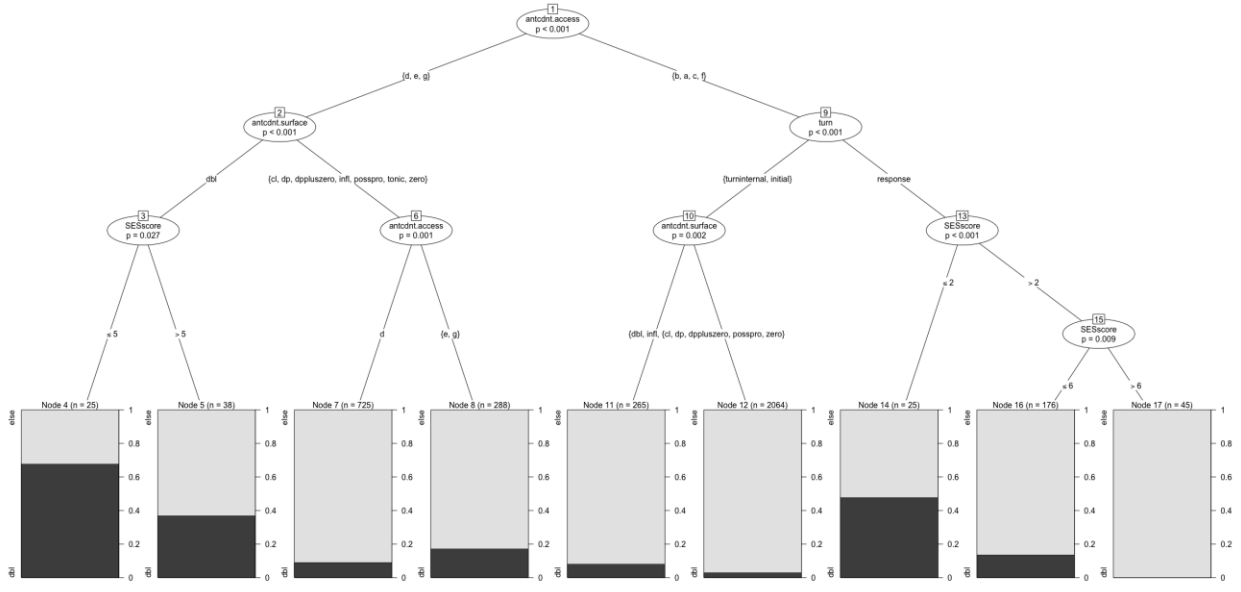
The results for SESSCORE in Model 4 support those from the previous analyses; participants with higher SESSCOREs use a reduced variety of subject forms (higher rates of SCs at the expense of X2SBJs and Ø subjects). A one point increase in SESSCORE was associated with a

disfavoring effect on the realization of X2SBJs (LOD = -0.20, $p < 0.01$), when compared with ALL ELSE.

For the factor CLAUSE CHAIN POSITION, targets contained in CHAIN INITIAL position, as opposed CHAIN INTERNAL clauses, were a favoring condition for X2SBJs (LOD = 0.65, $p < 0.01$), over ALL ELSE. Like the results found for question RESPONSEs, the result in which X2SBJs are favored in CHAIN INITIAL clauses is likely due to their being used contrastively, to reinforce the topic-status of a discourse referent, or to reintroduce an old referent.

To more closely inspect interactions among the predictors in Model 4, a conditional inference tree was grown and additional predicted probabilities were plotted. In the predicted probabilities plots, the effects of two or three constraints were examined (all others held constant). Figure 51 shows the conditional inference tree grown for all the predictors considered in Model 4.

Figure 51. Conditional inference tree for outcome X2SBJs over ALL ELSE (SCs and Ø subjects) for predictor terms SURFACE FORM OF THE ANTECEDENT, ANTECEDENT ACCESSIBILITY PATTERN, TURN, SESSCORE, and CLAUSE CHAIN POSITION, Analysis 4.

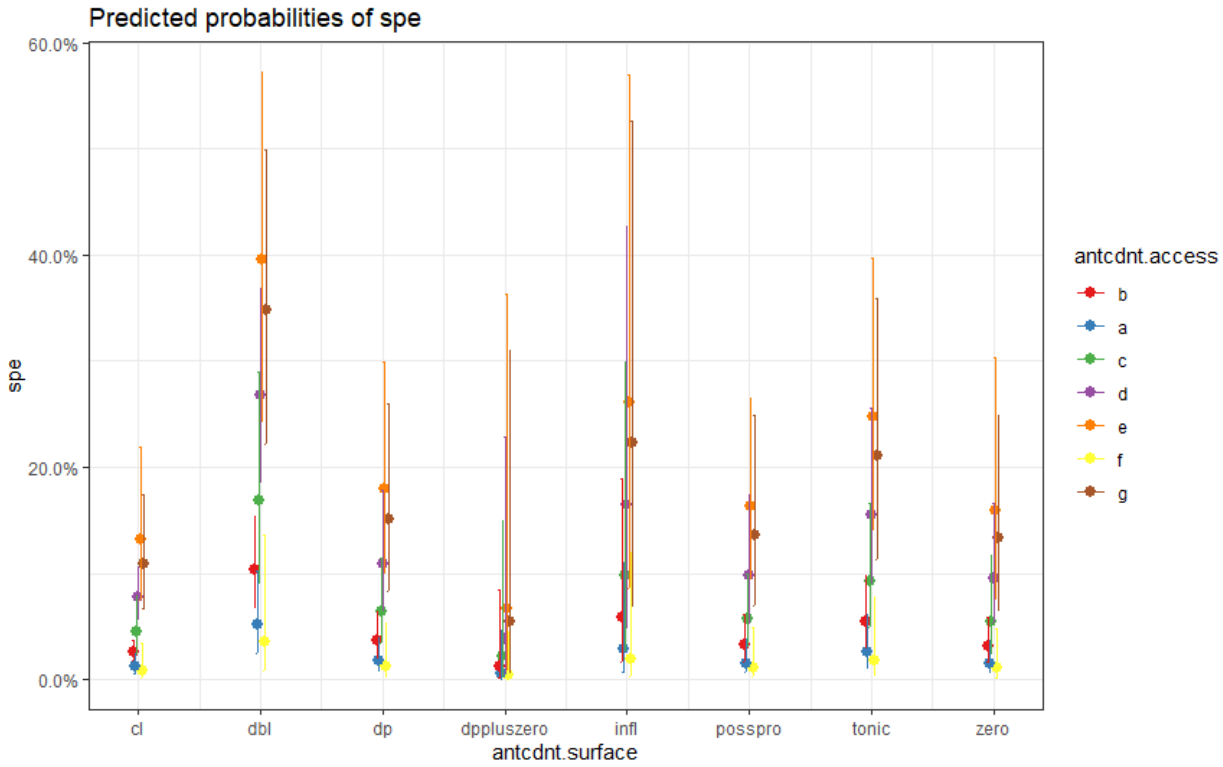


The first splitting branch reveals a significant difference for ANTECEDENT ACCESSIBILITY PATTERN ($p < 0.001$); contexts in which anaphoric targets and antecedents are in adjacent

IU/clauses (PATTERNs A, B, C, and F) pattern differently from contexts where targets and antecedents are in non-adjacent clauses (PATTERNs D, E, and G). These latter ANTECEDENT ACCESSIBILITY PATTERNs are on the left-branching side of the tree, where it can be seen that the SURFACE FORM OF THE ANTECEDENT becomes significant ($p < 0.001$); X2SBJ antecedents patterned apart from all other antecedent types. When the antecedent was a X2SBJ, the participants' SESSCORE became relevant ($p < 0.05$); participants with SESSCOREs ≤ 5 exhibited a stronger X2SBJ-to-X2SBJ priming effect (though this priming context was also highly favorable for realizing X2SBJs in the speech of participants whose SESSCORE was >5). This shows that X2SBJs are being used to reintroduce old discourse referents that had previously be introduced by a X2SBJ, and that speakers with higher SESSCOREs are more restrictive with respect to the constraints that promote non-clitic subject anaphora.

For all other antecedent types, PATTERN D contexts were distinguished from PATTERN E and G contexts ($p < 0.01$) (regardless of the speakers' SESSCORE). The latter contexts were associated with more X2SBJs. This indicates that, absent priming effects, X2SBJs are more likely to resume non-subject antecedents in non-adjacent clauses (PATTERN E) or subject antecedents from a previous 'discourse chunk' (PATTERN G, reintroducing an old discourse topic), than they are to resume a subject antecedent in a non-adjacent clause (PATTERN D) or subjects in an adjacent clause (PATTERNs A, B, C, and F). The predicted probabilities of realizing X2SBJs over ALL ELSE (SCs and Ø) as an effect of SURFACE FORM THE ANTECEDENT (x-axis) and ANTECEDENT ACCESSIBILITY PATTERN (plotted points and error lines) are shown in Figure 52. It can be seen that the most probable context for realizing a X2SBJ is when the antecedent is a X2SBJ, superstrate INFLECTION, or a TONIC pronoun in a non-adjacent clause.

Figure 52. Predicted probabilities of realizing X2SBJs over ALL ELSE (SCs and Ø subjects) as an effect of SURFACE FORM THE ANTECEDENT by ANTECEDENT ACCESSIBILITY PATTERN, mixed-effects binomial logistic regression, Model 4, Analysis 4.



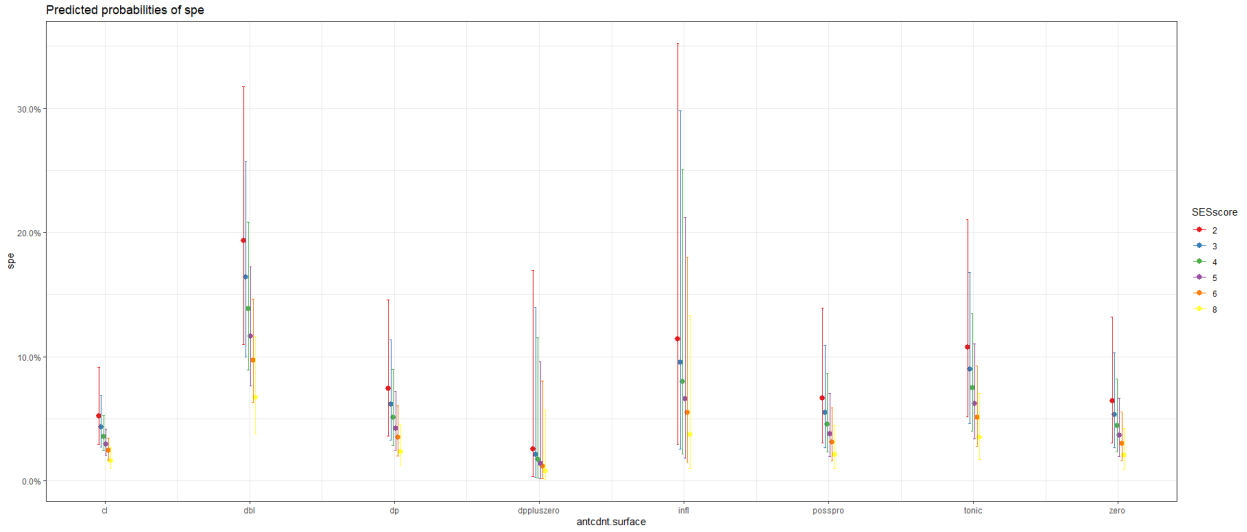
Returning to the conditional inference tree above (Figure #), for all ANTECEDENT ACCESSIBILITY PATTERNS in which antecedent and anaphor were in adjacent clauses, TURN taking became significant ($p < 0.001$). Targets in clauses that were TURN INTERNAL and TURN INITIAL (participant-initiated turns not prompted by a question), were distinguished from target anaphora contained in question RESPONSEs. In the former context, the SURFACE FORM OF THE ANTECEDENT was relevant ($p < 0.01$), in that X2SBJs, TONIC pronominals, and superstrate INFLECTION antecedents were more permissive of X2SBJ targets, as opposed to all other antecedent types.

In the RESPONSE context, SESSCORE again became significant, this time in two ways: the first relevant distinction is for an SESSCORE of two (the lowest score) ($p < 0.001$); participants with a score of two strongly associate targets in RESPONSE clauses with X2SBJs. There is yet another

SESSCORE-based distinction ($p < 0.01$); participants with scores >2 but ≤6 realized significantly more X2SBJs when compared to participants with a score >6. These results show that speakers with lower SESSCORES were more sensitive to the X2SBJ-to-X2SBJ priming effect and use more X2SBJs in question responses.

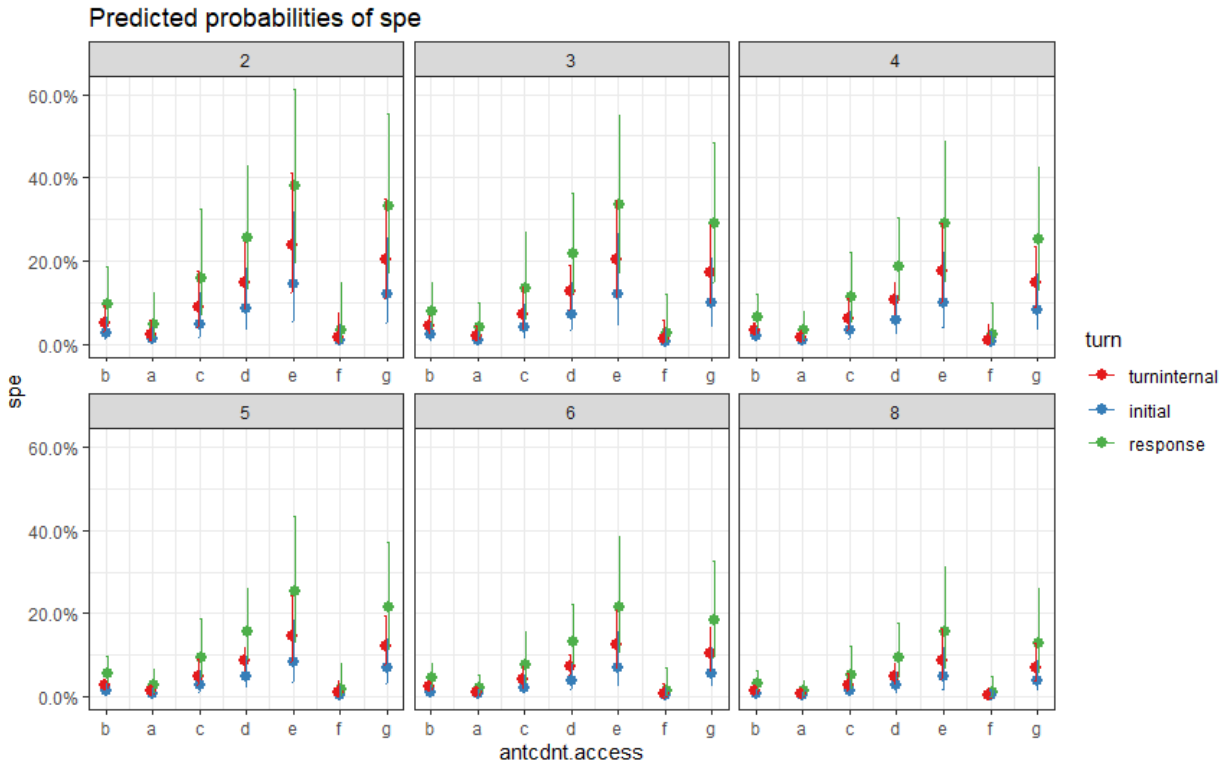
The predicted probabilities of realizing X2SBJs over ALL ELSE (SCs and Ø) as an effect of SESSCORE (x-axis in plot) and SURFACE FORM OF THE ANTECEDENT (plotted lines and error shading) are shown in Figure 53.

Figure 53. Predicted probabilities of realizing Ø subjects over OVERT subjects as an effect of SESSCORE by SURFACE FORM OF THE ANTECEDENT, mixed-effects binomial logistic regression, Model 4, Analysis 4.



The predicted probabilities in Figure # show that X2SBJs were most probable when the antecedent was another X2SBJ, a TONIC pronoun, or superstrate INFLECTION, and when the speaker has a lower SESSCORE. The predicted probabilities of realizing X2SBJs over ALL ELSE (SCs and Ø) as effect of ANTECEDENT ACCESSIBILITY PATTERN (x-axes in plots) and TURN (dots and error bars), with facetted panels for SESSCORE are shown in Figure 54. This plot shows that X2SBJs were most probable when the antecedent is in a non-adjacent clause, occurred in a question response, and the speaker had a low SESSCORE.

Figure 54. Predicted probabilities of realizing Ø subjects over OVERT subjects as an effect of ANTECEDENT ACCESSIBILITY PATTERN by TURN and SESSCORE, mixed-effects binomial logistic regression, Model 4, Analysis 4.

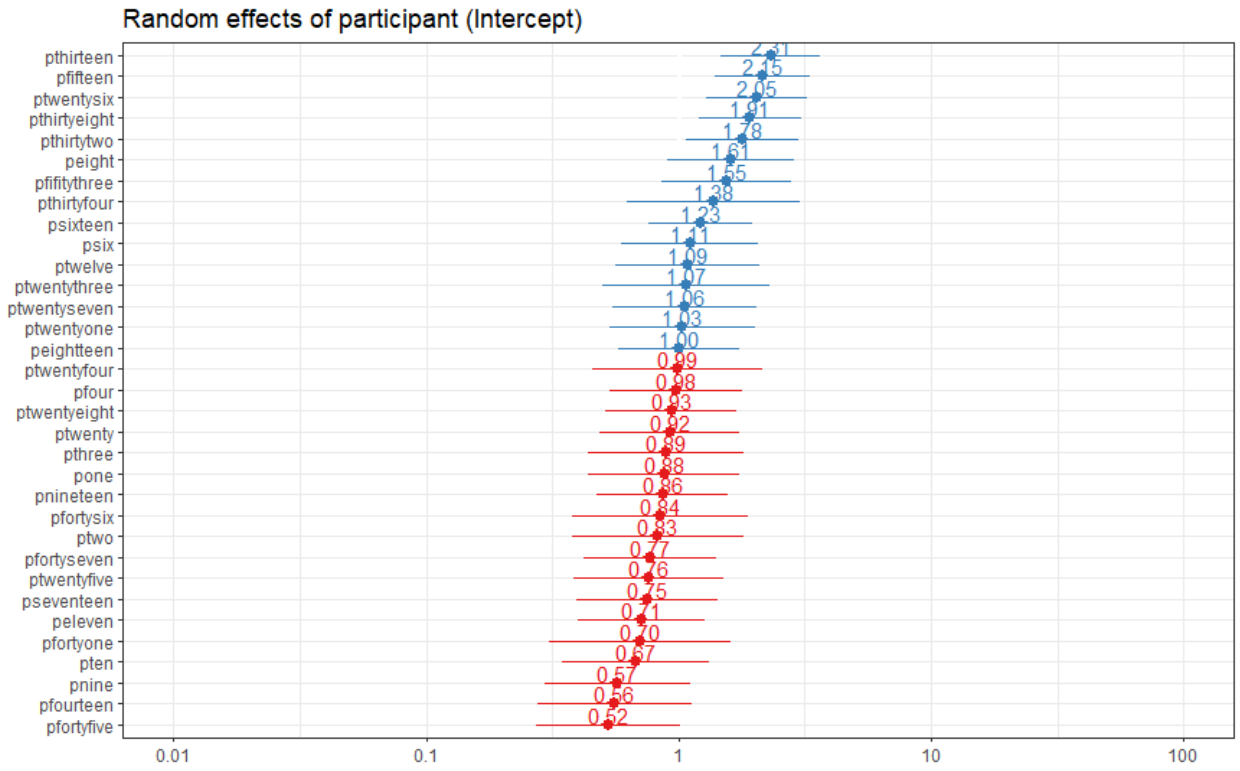


The results for the fixed-effects in Model 4 have revealed that X2SBJs are promoted by a priming effect and are favored when their antecedent is in a non-adjacent clause. They are used in responses to questions, particularly by speakers with lower SESSCORES, and also in chain-initial clauses; in all these contexts, X2SBJs are being used contrastively, to reinforce the topic-status of a discourse referent, or to reintroduce an old discourse referent. In continuation of the trend observed for SESSCORE, generally speaking, participants with higher scores use a reduced variety of subject anaphora, including less X2SBJs.

The caterpillar plot in Figure 55 displays the conditional modes (BLUPs) for each PARTICIPANT and the associated confidence intervals in Model 4. The errors bars (for confidence

intervals) that do not overlap with the intercept line (1) differ significantly from the population in patterns of SPE. The descriptive results for these subjects are briefly examined.

Figure 55. Plot of the random effects (BLUPs) for PARTICIPANT, mixed-effects binomial logistic regression, Model 4, Analysis 4.



At the upper bounds of the plot, speakers P13, P15, P26, P38, and P32 differed from the population average in realizing high rates of X2SBJs. Speaker P13 was 18 years old, their childhood dialect zone was SANTIAGU SENTRU, but their adulthood dialect zone was SATNIAGU SUL. Speaker P13 had an SESSCORE of 6 and realized 21 X2SBJs out of 159 tokens (=13.2%). Speaker P15 was 19 years old, from SATNIAGU SUL, had an SESSCORE of 6, and realized 24 X2SBJs out of 186 tokens (=12.9%). Speaker P26 was 24 years old, from DJARMAI, had an SESSCORE of 5, and realized 24 X2SBJs out of 110 tokens (=21.8%). Speaker P38 was 51 years old, from SANTIAGU SENTRU, had an SESSCORE of 4, and realized 21 X2SBJs out of 140 tokens (=15%). Finally, speaker P32 was 55 years old, from SANTIAGU SENTRU, had an SESSCORE of 6, and realized 15 X2SBJs out of 144 tokens

(=10.4%). The results for the random factor PARTICIPANT indicate that patterns of X2SBJ use may be sensitive to fluctuations stemming for idiosyncratic patterns of the individual speakers, their speech-style, or some other non-linguistic-structural related effect.

6.4 CHAPTER CLOSING REMARKS.

To summarize, the main results that obtained across all analyses in this chapter were:

i. All subject types are sensitive to structural priming.

ii. Ø subjects are associated with semantically referentially deficient lexical DP antecedents (i.e. inanimate, nonspecific, and indefinite reference).

   a. Partially as consequence of (ii) (but not entirely), Ø subjects are favored in the 3rd person (like in Brazilian Portuguese).

   b. Ø subjects may be favored in narrative style (rather than conversational style), but this may also be a consequence of (ii) stemming from the greater numbers of 3rd person referents and fewer overall number of discourse referents in the PICTURE DESCRIPTION NARRATIVE TASK.

   c. The effects in (i) and (iia-b) are activated by PROSODIC or BOTH prosodic and syntactic LINKING.

   d. Speakers with higher SESSCOREs are more sensitive to the constraints in (i), (ii.a), and (ii.c).

iii. Ø subjects are disfavored in embedded contexts, except *pa*-clauses (like in Papiamentu).

iv. Ø subjects are disfavored when there is a SWITCH in TMA FRAME between clauses, presumably because TMA SWITCHes decrease referential continuity and antecedent accessibility.

v. X2BJs are favored when their antecedent in a non-adjacent clause. This is because they are switch-reference devices and are used to introduce old discourse antecedents.

vi. X2BJs are favored in question RESPONSEs and CHAIN INITIAL clauses, where they serve to contrast with another discourse referents, enhance the topicality of a referent, or reintroduce an old discourse antecedents.

vii. Speakers with higher SESSCOREs generally use less of a variety of subject forms and more SCs because they are more restrictive with respect to the contexts in which non-clitic forms are most likely to occur. Speakers with lower SESSCOREs are less sensitive to all of the above constraints.

Having presented the descriptive and inferential results from four models of SPE in a corpus of CVC, I turn now to the discussion/conclusions section where these results are discussed in light of their significance for the relevant questions in this thesis related to the status of each subject elements under examination (X2SBJs, SCs, and Ø). The conclusions drawn are then used to reflect on formal cross-linguistic models of SPE (the NSP, the Typology of Structural Deficiency, functionalist and 'cognitive linguistic' typological models) and in light of SPE in other Romance-origin vernaculars and languages with tonic & atonic subject pronominal inventories and (discourse-oriented) Ø subjects.

Chapter 7: Discussion and conclusions

This study has explored the topic of Subject Pronoun Expression (SPE) in Cabo-Verdean Creole (CVC), a Portuguese-based contact vernacular spoken in the Republic of Cabo Verde. It is the first study to explore this topic by combining empiricist/quantitative, rationalist/formalist, and 'Probabilistic Grammar' (Claes 2017) perspectives. This trans-paradigmatic approach applied concepts of hierarchical structure and formal categories from the Generative Grammar (GG) tradition in the design of several predictors that formed the basis for a variationist sociolinguistic analysis, while also borrowing explanatory frameworks from functionalist, usage-based, 'cognitive linguistic', and typology-oriented approaches. The motivation for this study was to address unresolved debates over the morphosyntactic status of nominative anaphoric elements like subject clitics (SCs), double-subject constructions (tonic pronoun + subject clitic; X2SBJ), and null/zero subjects (Ø), the properties of which have been at the center of protracted disputes in several languages and across theoretical paradigms.

By adopting a quantitative methodology and considering diverse theoretical perspectives, I was able to provide the first delimitation of the envelope of variation or the variable context for SPE in CVC. In order to account for the three-way dependent response variable associated with nominative pronominal anaphora in the CVC subject domain (e.g. SCs, X2SBJs, and anaphoric Ø), a still underused procedure in linguistics – multinomial logistic regression (MLR) – was adopted as one of the methods for analyzing the dataset. The descriptive and inferential results allowed me to compare the morphosyntactic and discursive distribution of the three subject anaphora as they correlate with several predictor factors related to anaphoric relationships with discourse antecedents, clausal organization, domain-general cognitive processes like structural priming, the formal semantic properties of 'referential deficiency', and the influence of social categories like those related to the speakers' Socioeconomic Status (SESSCORE).

Chapter 1 was an introductory chapter; it presented the main themes of the dissertation (Section 1.1), established the theoretical and methodological framework to be applied (Section 1.2), and provided a summary of the components of this thesis (Section 1.3).

Chapter 2 served as a synopsis of the language-external history of Cabo Verde with attention to the sociocultural context in which CVC developed and its relationship to other languages in the Upper Guinea Coast family of Portuguese-lexifier creoles. I began by noting that, to the best of our knowledge based on archeological evidence, Cabo Verde was uninhabited prior to settlement of Santiago by the Portuguese, Genovese, and Venetians. Quickly thereafter, enslaved people of Wolof, Mandinka, Fula, and Balanta descent, among many other societies/ethnic groups from the adjacent Upper Guinea Coast, were brought to Santiago, where many of them were forced to undergo a process of *ladinização*. This process entailed forced conversion to Catholicism, forced assignment of a specific skill or trade, and forced instruction on the rudiments of Portuguese. Most of the enslaved people who suffered *ladinização* were later sent to the Iberian Peninsula or the Americas, where they would be sold at a higher price than if they had been taken directly from the Upper Guinea Coast as so-called '*boçales*' (see Section 2.1).

Those who remained in Santiago either came to reside in the urban center of Ribeira Grande, where they labored as artisans and in other urban trades, or they toiled in the *latifúndios* 'plantations' of the interior. European-descended landowners, navigators, merchants, government officials, bureaucrats, and clergy, came to be drastically outnumbered by free and enslaved people of mixed-race or African descent, creating sharp demographic imbalances. By the middle of the 16th century, Santiago had already become a thoroughly 'creole' society (Green 2006:84-86), and as early at the turn of that century, CVC had already become mostly consolidated as a code distinct from its contributing languages (Quint 2000a,b; Jacobs & Quint 2016).

After the middle of the 16th century, Santiago's near monopoly on maritime trade with the adjacent Upper Guinea Coast started to be challenged by other European maritime powers (see Section 2.3). These challenges, combined with natural disasters like cycles of drought and famine, and internal political disputes, all began to slowly chip away at Santiago's prosperity and

geopolitical prominence. By the end of the first quarter of the 17th century, Santiago had slipped into a prolonged period of economic stagnation and decline. During that time, the mixed-race 'creole' character of the people, society, and culture, became reinforced, and the agriculturally oriented population of the rural mountainous interior continued to grow (Green 2006:242-245).

This developmental trajectory might be viewed as a historical 'prototype' because many other plantation and entrepôt colonies throughout the Atlantic - where contact vernaculars historically labelled 'creoles' repeatedly emerged - followed a similar cycle of boom-and-bust socioeconomic growth and collapse contingent on trans-Atlantic slavery. This cycle was accompanied by correspondingly violent demographic shifts and imbalances, and the emergence of a syncretic mixed-race culture speaking a new vernacular characterized by contact-induced morphosyntactic restructuring (Sections 2.1-2.3).

Following Parkvall (2000), Jacobs (2010, 2012), and Jacobs & Quint (2016), it was concluded, based on a combination of sociohistorical, linguistic, and population-genetic evidence (Verdu *et al.* 2017), that the Santiago variety of CVC must have become mostly consolidated in its 'proto-'state somewhere between the last quarter of the 15th and the turn of the 16th century; this proto-vernacular was dispersed throughout the region, taken to other Cabo-Verdean islands, the adjacent mainland, and even across the Atlantic to locales such as Cartagena de Índias, Maranhão, and eventually Curação (Jacobs 2012). In this way, the proto-vernacular that would eventually become the Santiago variety of CVC (Badiu), was the source variety for the other Portuguese-lexifier contact vernaculars of the Upper Guinea Coast group: Guinea-Bissau Creole, Ziguinchor Creole, Papiamentu, and the varieties of CVC spoken on other islands of the archipelago (Quint 2000a; Jacobs 2012; Jacobs & Quint 2016) (Section 2.3).

The remained of Chapter 2 was dedicated to the long period of stagnation in CVC from the 17th century to the late abolition of slavery in the last quarter of the 19th century (Section 2.4), the post-slavery period and the end of the colonial era (Section 2.5), and a sociolinguistic sketch of modern-day Cabo Verde with attention to the social status of CVC and issues of diglossia (Section 2.6).

Chapter 3 was dedicated to a summary of a historical-reconstruction of the subject pronoun system presented in Quint (2000a) and Lang (2012) (Section 3,1), a brief review of the major subject domain and subject-domain-adjacent morphosyntactic properties of Late Medieval/Early Classical Portuguese (Section 3.2), and an overview of the subject pronoun systems of two of the major substrate languages involved in the formation of CVC: Wolof and Mandinka (Section 3.3).

Lang's (2012) reconstruction, and Quint's (2000a) prior grammar, proposed that CVC tonic subject pronoun developed in part from the tonic personal pronouns of Late-Medieval/Early Classical Portuguese; (atonic) proclitic subject pronouns developed in part from transformations in the tonic subject pronoun inventory (likely by way of grammaticalization), and in part from cross-linguistic convergence between morphophonemes from the superstrate and substrates that produced the 1st person SCs (='person markers'). Lang (2012) argues that these 1st person SCs, in conjunction with their tonic pronoun counterparts, triggered an innovation by analogy in the 2nd person that caused both the tonic and atonic forms to lose their coda *-s*.

After the review of Lang's (2012) reconstruction, a brief summary of the morphosyntactic properties of Late-Medieval/Early Classical Portuguese highlighted the verb-second (V-2) features of that language and its *pro*-drop status as a consistent Null Subject Language. Both of these features must have been lost very early in the morphosyntactic restructuring processes in the years following the initial settlement of Santiago in the mid-15th century.

A review of the subject pronoun inventories of Wolof and Mandinka ended Chapter 3. It was observed that Wolof subject markers undergo a complex series of transformation based on their morphosyntactic, discursive, and semantic/pragmatic role in the clause, as well as tense/mood relationships and the status of their host, and that these complex transformations are simply absent from the properties of CVC SCs (Section 3.3.1). Nevertheless, it was taken to be unsurprising that CVC developed a bipartite subject pronoun system with a tonic-atonic opposition, since both its primary substrate languages have such disjunctive subject pronominal inventories. This development may also be an example of cross-linguistic convergence, since it is well known that such bipartite subject pronominal systems resulted from diachronic changes in several Romance

varieties, and since the disjunctive mode of pronominal organization also existed among the superstrate's oblique pronouns. Further, it was noted that Zribi-Hertz & Diagne (2002) provide an analysis of Wolof person-markers in which these atonic nominative anaphora were shown to lead a 'double-life' as genuine pronouns and inflectional affixes. This is essentially analogous to the 'ambiguous anaphoric person marker' analysis adopted by Siewierska (2004), a point that I will revisit later in this chapter.

In Chapter 4, I reviewed the previous literature on SPE and anaphora resolution in CVC and related languages. I began by outlining the various ways of categorizing overt subject pronouns in CVC (Section 4.1); I labelled these the 'bipartite' and 'tripartite' analyses. I also provided numerous examples of the distributional patterns associated with SCs and X2SBJs, as they have been explored in Pratas (2004) and Baptista (2002). Since much of the prior research on SPE in CVC has been developed using the methodological tools of GG or the Minimalist Program, I then turned to a review of analyses of CVC overt pronominals that have used these frameworks (Sections 4.3 and 4.5.2). These previous analyses draw on formal models of cross-linguistic variation in SPE such as the Null Subject Parameter and the Typology of Structural Deficiency. Particular consideration was given other Lusophone- and Iberian-origin Romance vernaculars (Section 4.5.1), to other languages with tonic-atonic oppositions in their subject pronominal inventory such as Gallo-Romance and Northern Italian Dialects (Section 4.2), to partial Null Subject Languages (Section 4.5.1-4.5.2), and languages that make use of ∅ anaphora despite lacking extensive person-number inflection (Section 4.5.3).

The previous literature on SPE in CVC in the GG tradition centers on two interrelated issues: the morphosyntactic status and underlying structural position of SCs and the availability and nature of ∅ subjects. With respect to this first issue, proponents of the 'inflectional affix hypothesis' (following the terminology in Culbertson 2010 on SCs in Colloquial French and cross-linguistically) has argued that SCs are bound inflectional affixes, akin to person-number agreement morphology, attributing them the status of syntactic heads rather than independent pronouns receiving nominative case in a argumental (specifier) position (cf. Baptista 1995; 2002).

Proponents of the 'phonological clitic hypothesis' have countered that SCs are indeed genuine independent subject pronominals occupying the canonical subject position, and that cliticization merely occurs at the phonological level (Pratas 2004; Costa & Pratas 2008, 2013).

While the above accounts disagree as to the hierarchical clausal position of SCs in CVC, we saw in Chapter 4 that Baptista (2002) and Pratas (2004) nonetheless converge in adopting the Typology of Structural Deficiency (Cardinaletti & Starke 1994, 1996, 1999) as a means of distinguishing each element in the CVC subject domain according to a set of morphological, syntactic distributional, prosodic, and referential-semantic properties. This typological model posits that there are three universally underlying classes of syntactic element - strong, weak, and clitic - and each of these exhibit a set of properties corresponding to the degree to which they are 'structurally deficient'. For the present study, the most of important of the properties were those related to the referential semantics of each of these forms[84].

The claim is that strong forms are fully referentially specified and consequentially will tend to be [+animate, +definite, +specific]. These properties are asymmetrical and overlapping among the three classes. Deficient forms (weak and clitic), due to their 'referential deficiency', receive specification from a coreferential antecedent. Deficient forms can be [-animate, -deficient, -specific], though they can assume the opposite values as well. The morpheme thought to be underlying null subjects, *pro*, is assumed to be a deficient pronoun that exhibits the referential semantic properties associated with its class (see Section 4.3). Independent empirical support for this has come from the study of SPE in Brazilian Portuguese, in which ∅ is probabilistically promoted when its antecedent bears nonspecific or inanimate reference, leading Kato & Duarte (2003, 2005) and Duarte & Soares da Silva (2016) to posit an "avoid referentially deficient pronoun constraint" as a modification of Montalbetti's (1984) Overt Pronoun Constraint (see Sections 4.5.1-4.5.2).

84 This is because classic morphosyntactic tests for clitichood have proven incapable of establishing a clear-cut distinction between independent pronominal clitics and bound affixes. Further, a full investigation of the prosodic properties of each element were beyond the scope of the present study.

Chapter 4 also explored how controversies in other languages with tonic and clitic subject pronouns have been characterized by similar disagreements; these include French, Krèyol Ayisen, Northern Italian Dialects, other Indo-European languages (cf. DeGraff 1993, Cardinaletti & Starke 1999; Poletto 2000; Culbertson 2010, Poletto & Tortora 2016; *inter alia*) (Section 4.2), languages from all four 'phyla' on the African continent (Bresnan & Mchombo 1987; Kari 2017), and in indigenous languages from various families throughout the Americas and Oceania (cf. Givón 1976, 2001[1984], 2012, 2017; Siewierska 2004) (Section 4.4).

The debate in the GG tradition is reflected by a parallel debate in the 'Probabilistic Grammar' paradigms (Claes 2017): the morphosyntactic status of 'person markers' (~SCs) is disputed; some have argued that they are pronominal agreement markers that engage in nonlocal anaphoric agreement (~phonological clitics, i.e. genuine subject pronouns), others have argued they are syntactic agreement markers that engage in local grammatical agreement (~inflection affixes, i.e. syntactic heads) (cf. Bresnan & Mchombo 1987). Others argue that many languages have ambiguous person agreement markers capable of engaging in nonlocal anaphoric agreement or local grammatical agreement, often depending on the properties of the controller and antecedent accessibility (Siewierska 1999, 2004) (Section 4.4).

Chapter 4 considered how disagreements over the status of SCs in CVC has led to debates over the status of CVC under the Null Subject Parameter (cf. Baptsita 1995, 2002 vs. Pratas 2005; Costa & Pratas 2008, 2013) (Section 4.5.2). Under inflectional affix hypothesis SCs are heads in INFL, this would imply that, in clauses with a lone SC, the specifier position must necessarily be analyzed as empty, presumably filled by *pro*. In X2SBJ constructions in which a tonic pronoun or lexical DP doubles the SC, the inflectional affix hypothesis would situate the tonic pronoun/lexical DP in the specifier or canonical subject position. Given the predominance of clauses with only a singleton SC in CVC, the inflectional affix hypothesis ultimately leads to a classification of CVC as a fully Null Subject Language under the Null Subject Parameter (Baptsita 1995, 2002).

On the other hand, under the phonological clitic hypothesis, SCs occupy the canonical subject position, thus *pro* is understood to be banned in root contexts and is thought only to emerge

in embedded contexts when c-commanded by a limited set of operators (Costa & Pratas 2013). This implies that X2SBJ constructions involve a left-peripheral tonic pronoun/lexical DP (Pratas 2004), and that CVC is at best a partial Null Subject Language, since ∅ subjects would only occur with weather and existential/presentational predicates, impersonal or generic constructions, or in the aforementioned embedded contexts (Pratas 2004; Costa & Pratas 2008, 2013).

Chapter 4 also explored theoretical paradigms that operate outside the GG or Principles & Parameters framework (within which the Null Subject Parameter was developed). Under Probabilistic Grammar approaches (see Section 4.5.3) - which include variationist sociolinguistics, functionalism, language typology, and usage-based and 'cognitive linguistic' approaches (Claes 2017) - formal devices such as *pro* or PRO are not required, the ability for a language to deploy ∅ is not associated with a 'cluster of properties', and ∅ does not rely on some formal identification mechanism. Instead, it is assumed that ∅ anaphora are promoted by a collusion of predictors; these include language-specific constraints, domain-general cognitive processing constraints, constraints related to antecedent-anaphor relations, the semantics of discourse referents, and extra-linguistic factors. The latter may be tied to individual-specific social categories (e.g. social group or class, occupation, age etc.) or to non-individual-specific language-external factors like the data collection procedures/materials/location, the identity of the interviewer or researcher, or other situational communicative concerns (Li & Thompson 1976; Givón 1976, 1983b; 2001[1984], 2017; Ariel 1990, 2001; Duarte 1993, 1995; Barbosa, Kato & Duarte 2005; Bullock & Toribio 2009; Martínez Sanz 2011; Otheguy & Zentella 2012; Torres Cacoullos & Travis 2014, 2015, 2019; Carvalho, Orozco, & Lapidus Shin eds. 2015; Duarte & Soares da Silva 2016; Speyer 2016; Wagner 2016; Claes 2017; *inter alia*).

Another important point highlighted in Chapter 4 was that the notion of Ø subjects as "typological exotica" has been rejected (Givón 2017) (Section 4.5.3). Many languages, including those usually considered non-Null Subject Languages, or those that lack 'rich' inflectional morphology, nonetheless still make use of Ø anaphora. This is the case to the extent that some

researchers have asserted that the majority of the world's languages are partial Null Subject Languages (Vainikka & Levy 1999; Wratil 2011).

Both ∅ anaphora and atonic SCs have also been found to be favored (and to compete with each other) in contexts of high antecedent accessibility[85]; this involves an overlapping of numerous conditions, among these: adjacency in the antecedent's and anaphor's clauses/Intonational Units, prosodic and syntactic linking, the anaphor's clause-chain position, the morphological composition and syntactic role of the antecedent, the anaphoric distance between antecedent and anaphor, the semantic-referential properties of the antecedent, clause type and clausal embedding, the lexical semantics of the verb, and continuity in the Tense-Mood-Aspect 'frame' across clauses, among others (cf. Givón 1976, 1983b, 2017; Ariel 1990, 2001; Martínez Sanz 2011; Otheguy & Zentella 2012; Carvalho, Orozco, & Lapidus Shin eds. 2015; Duarte & Soares da Silva 2016; Claes 2017; Torres Cacoullos & Travis 2019; *inter alia*).

Chapter 5 addressed the methodology adopted in the present study. I first recapitulated the theoretical and methodological frameworks (Section 5.1), and the broad themes of the dissertation (Section 5.2). Specific research questions and hypotheses were offered for each predictor variable as they relate to the three anaphoric subject elements in CVC (Section 5.3), and after that, the envelope of variation for SPE in CVC was delimited based on a range of examples from my corpus (Section 5.4). I also discussed the data collection materials, speakers in the sample, and locales where speech was collected (Section 5.6), and considered research questions and hypotheses related to the language-external factors (Section 5.3.2.1). The coding procedure and levels of the predictor factors were specified (Section 5.3), and in the last section of Chapter 5 (Section 5.7), I introduced the statistical procedures used in the present study, which were then elaborated further in the following chapter.

---

[85] Antecedent accessibility is also related to overall antecedent activation as an effect of its 'salience' in the discourse and the referential continuity between antecedent and anaphor. There are many ways to encode these properties in a quantitative study, only some of which have been included in the present dissertation. A comprehensive exploration of the many aspects of discourse referent salience, activation, and accessibility has been beyond the scope of this thesis; but see Holler & Suckow (2016) and Speyer (2016) for a detailed list of references with competing views on what properties comprise accessibility/salience/activation/continuity with respect to anaphora.

In the previous chapter, Chapter 6, the results were presented from four analyses in order to explore variability in the realization of three subject forms in CVC: SCs, X2SBJs, and anaphoric Ø. Chapter 6 was intended as a technical description of the results, while in this chapter the results will be discussed more generally with respect to their relevance for the broad themes of this dissertation. Analysis 1 was exploratory; it was the first attempt at delimiting the variable context for SPE in CVC to be used in following analyses. In both the descriptive and inferential results, each subject realization was taken to be a possible SPE outcome; descriptive results looked at the distribution of all three forms, while in the inferential analysis, SCs were the baseline for comparison against the other two forms (Section 6.1).

Descriptive and inferential results from Analysis 1 were then used to delimit the envelope of variation for Analysis 2. Several contexts of non-variability were observed and refinements were made to the criteria for admitting potential Ø subjects into the envelope of variation. Analysis 2 also employed a multinomial logistic regression (MLR, i.e. three-way response variable). Another improvement to MLR 2 in Analysis 2 was that it did not suffer from the high degree of multicollinearity that MLR 1 in Analysis 1 did. Thus, results from Analysis 2 directly contribute to the discussion and conclusion drawn in the current chapter. Analysis 2 did not, however, consider the variance associated with random effects (Section 6.2). In order to account for such variance, and in order to confirm the results in Analysis 2, two binomial models were constructed; one of these, Model 3, compared Ø subjects against all other subject types (Section 6.3), while the other, Model 4, compared X2SBJ constructions again all other subject types (Ø and SCs) (Section 6.4). Discussion of the results in this chapter will consider the totality of the results and draw conclusions based on generalizations informed by their aggregate.

In the remained of this chapter, I summarize the results from Chapter 6 and discuss their broader implications. The first half deals specifically with the results for Ø subjects and X2BJs. Then, I consider what those results can tell us about the status of SCs in CVC and cross-linguistically, which leads me to some observations about how CVC might be situated in light of typological models concerned with classifying languages based on pronominal expression in the

subject domain (the Null Subject Parameter, the Typology of Structural Deficiency, diachronic and typological perspectives and atonic subject pronouns and zero anaphora).

7.1 DISCUSSION OF THE RESULTS FOR NULL SUBJECT/ANAPHORIC ZERO AND DOUBLE-SUBJECT CONSTRUCTIONS.

In the first half of this chapter, I discuss the results found for Ø subjects and X2BJs. The first subsection addresses overall rates of these forms. The second discusses the results for structural priming, antecedent accessibility, and referential deficiency. The third subsection deals with the results for constraints related to clausal organization, anaphoric relationships across clauses and the discourse, clause type, TMA frame, and turn-taking. In the second half of the chapter, I discuss the results for the language-external factor SESscore and consider what the overall results can tell us about the status of SCs in CVC and cross-linguistically.

7.1.1 The overall distribution of subject forms.

Overall rates of each subject pronominal form revealed SCs to be by far the most frequent subject type (~81%), with Ø subjects occurring at rate of about 12% in the broad-envelope analysis (Analysis 1) and about 11% in the restricted-envelope analyses (2 and 3). X2SBJs occurred at just under 7% in Analysis 1 and just above 7% in Analyses 2 and 4. It was noted that rates of Ø subjects were close to what Duarte & Soares da Silva (2016) found for 2nd person subjects in Brazilian Portuguese (BP) (~10%), but much lower than the 41% found for 3rd person referents in BP (~41%). On the other hand, the overall rate of X2SBJs in CVC was comparable to rates of overt (tonic) subject pronouns in the 3rd person for Italian, this language being a consistent NSL by all accounts. With respect to this latter observation, this might be taken as indicative that X2SBJs in CVC (and possibly other languages with tonic-atonic amalgamations) likely serve many of the same discourse/pragmatic functions as singleton tonic pronouns in languages that do not rely on X2SBJs.

However, overall rates of overt/null SPE have also been described as "equivocal" (Torres Cacoullos & Travis 2019:2), since they vary greatly across languages (even those considered to be consistent *pro*-drop languages), across varieties of the same language, and even across contexts (morphosyntactic and communicative/situational) in the same variety (Wratil 2011; Posio 2012; Torres Cacoullos & Travis 2019; *inter alia*). What is more, rates of SPE reveal little about the variable-rule system underlying anaphora resolution. As such, the present study places much greater emphasis on the probabilistic distribution of the effects of independent predictors on SPE outcomes so as to extrapolate the findings for CVC to be able to draw conclusions about cross-linguistic patterns of SPE in languages with disjunctive tonic-atonic subject pronominal systems.

7.1.2 Priming, antecedent accessibility, and 'referential deficiency'.

In all the analyses presented, SURFACE FORM OF THE ANTECEDENT was the most important predictor variable. Two effects were associated with this factor: one was linguistic structural priming (for both X2SBJs and Ø), the other one impacted Ø subjects and was related to the semantics of referents introduced by LEXICAL DPs and the DP + INTERVENING MATERIAL + Ø construction. The Ø-to-Ø priming effect, and the relationship between anaphoric Ø and 'referentially deficient' antecedents, were activated by LINKING between adjacent Intonational Units (IUs) containing antecedent and anaphor. These relationships were also contingent on anaphoric distance (recency), though as an independent predictor ANAPHORIC DISTANCE did not achieve statistical significance in MLR 2, and was not retained as a parameter following the model building procedure for any of the binomial regression models (Models 3 and 4). Nevertheless, Ø subjects only emerged with recent, anaphorically proximate antecedents, while X2SBJs emerged at longer distances, since the latter are used to reintroduce old discourse antecedents.

With respect to priming, X2SBJs and lone TONIC pronouns were found to prime subsequent X2SBJs, while Ø was primed by Ø antecedents. In the case of the former relationship, the priming effect usually obtained across non-adjacent clauses/IUs. These results indicate that

X2SBJs have an equivalent discursive status to singleton tonic pronouns, at least in so far as their priming values are concerned. This is unsurprising since X2SBJs are the primary way to deploy a tonic pronoun in CVC with all verbs except copular *e* (which does not admit SCs).

The Ø-to-Ø priming effect often occurred in Ø-chains with short anaphoric distances between anaphor and antecedent (147).

(147) *ali dj=E$_i$ bai embora, dja Ø$_i$ kansa di pega sapu*
here TMA=3.SG.CL$_i$ go away TMA Ø$_i$ tire of catch toad
i dja Ø$_i$ ba si kaminhu
CONJ TMA Ø$_i$ go 3S.POSS path
'Now here he goes away, [he] tired of catching toads and [he] went on his way.'

X2SBJs were only used at short distances contrastively, to switch back-and-forth between discourse referents (148). The priming effect obtained at longer anaphoric distances when the X2SBJ served TO reintroduce an old discourse referent that had previously been born by a X2SBJ or TONIC antecedent.

(148) *Mi=N ta dizeja pa el=e ser un.. algên..*
1.SG.WK=1.SG.CL TMA desire COMP 3.SG.WK=3.SG.CL COP DET … someone…
un ómi di manhã
DET man of tomorrow
'I hope for him to be a man… a man of tomorrow.'

Linguistic structural priming has repeatedly been shown to exert a powerful psychological effect on the repetition of linguistic forms (Bock 1986; Bock & Lobell 1990; Lobell & Bock 2003; *inter alia*), not just in the laboratory setting, but also in everyday conversational discourse (Weiner & Labov 1983; *inter alia*.), as a catalyst for cross-linguistic influence in contexts of

language contact (Poplack 1980; Lobell & Bock 2003; Meijer & Fox Tree 2003; *inter alia*), in corpora (Szmrecsanyi 2005; Schwenter 2017), as a mechanism of "implicit learning" or other strategies in processing and language acquisition (Onishi *et al.* 2002; Huttenlocher *et al.* 2004; Goldberg 2006; *inter alia*), and as an active constraint for variable SPE in Spanish (Cameron 1993; Flores-Ferrán 2002; Cameron & Flores-Ferrán 2004; Travis 2007; Claes 2017; *inter alia*), Portuguese (Bouchard 2018), English (Wagner 2016), and for variable SPE in these languages in contact with each other (Carvalho & Child 2011; Carvalho & Bessett 2015; Abreu 2012; Torres Cacoullos & Travis 2015, 2016, 2018; *inter alia*) (see Pickering & Ferreira 2008 for an article-length review of research into priming across subfields of linguistics).

Priming and antecedent accessibility appear to be a two-way street: as morphemes or morphosyntactic configurations become activated in the discourse, they are more likely to be repeated in subsequent congruent contexts where they are eligible variants; conversely, referents that are more 'activated'[86] are most likely to induce priming (the same morpheme or morphosyntactic configuration is more likely to be repeated). As Claes (2017:4) describes:

> structural priming […] can be accounted for as a residual activation effect: once a particular representation has been visited, it remains more activated than others for a period of time, giving it a head start over its competitors. At the same time, structural priming also appears to be a mechanism of implicit learning, which permanently adapts the ease of activation of constructions to observed patterns of usage (e.g. Goldberg 2006: 120–125; Pickering & Ferreira 2008: 447).

Structural priming is a "mechanical effect" (Travis & Torres Cacoullos 2012) or a "domain-general cognitive constraint" (Claes 2017), as it is related to working memory and other aspects of non-declarative memory (Bock *et al.* 1992; Chang *et al.* 2000; Han & Yim 2015; Heyselaar, *et al.* 2017; *inter alia*). Therefore, the priming effect is not specific to the language faculty; rather, the same cognitive processes that drive priming (e.g. working memory, implicit learning) are active in non-linguistic cognitive domains as well. Nevertheless, since priming has repeatedly been found to be among the most relevant constraints for SPE across languages,

[86] As result of other properties like their semantic value or by virtue of being anaphorically proximate, etc.

regardless of *pro*-drop status, it should be a major component of any model that attempts to offer a typologically and cognitively comprehensive account of patterns of SPE. One example of such a model was recently outlined in Claes' (2017) call for merging methods and insights from Probabilistic Grammar (variationist-sociolinguistics and corpus linguistics) and Cognitive Linguistics (adjacent to usage-based or functionalist traditions). This fusion combines the descriptive power of the former with the explanatory/theoretical strengths of the latter. In GG-related cognitive models that often adopt a modular architecture of the mind, there have already been several attempts at explaining interfaces between the language faculty processing center and language-external cognitive domains (cf. Jackendoff 2015). It would not require much alteration to these models to account for processes driven by implicit learning, working memory, and activation, such as priming (and anaphoric distance).

For SPE in CVC specifically, priming is also manifested in the triggering effect of Ø antecedents on subsequent Ø targets, observable at short anaphoric distances (mostly up to 7 words, seldom beyond 10). Yet, another set of results for ANTECEDENT SURFACE FORM indicated that the antecedent types LEXICAL DP and the DP + INTERVENING MATERIAL + Ø were found promote the realization of anaphoric Ø, particularly when these antecedents bore [-SPECIFIC], [-DEFINITE], or [-ANIMATE] reference. This selectional effect also obtained at short anaphoric distances, but it is clearly not a priming effect since it involved a change in surface morphological form. Further, accessibility hierarchies that rank referring expressions by salience tend to associate semantic properties such as non-specificity, indefiniteness, and inanimacy with reduced discourse salience, thus impeding accessibility. When an antecedent's referent is less salient, this usually reduces the probability of selecting anaphoric Ø over an overt form (cf. Givón 1983a:10; Ariel 1990:74; Siewierska 2004:149). For instance, Givón (2017:2) ranks indefinite NPs second lowest in his accessibility hierarchy based on referential salience and the morphosyntactic status of the antecedent, they are associated with the least referential continuity (example 100, above; repeated here as 149).

(149) Most common referential coherence devices

highest continuity

a.) zero anaphora

b.) unstressed anaphoric pronouns

c.) stressed independent pronouns

d.) definite NPs

e.) indefinite NPs

f.) modified NPs

lowest continuity

Furthermore, cross-linguistically, anaphoric Ø is usually more frequent after 1st and 2nd person antecedents (Speech Act Participants), since SAPs are highly 'topic-worthy' and facilitate antecedent accessibility, thus promoting the use of Ø anaphora (cf. Wratil 2011). But in CVC (like in BP), Ø subjects are favored over overt forms in the 3rd person and with semantically referentially deficient antecedents (cf. Duarte & Soares da Silva 2016). The relationship between referentially deficient antecedents and Ø in Brazilian Portuguese (BP) led Kato & Duarte (2003, 2005) to propose an 'avoid referentially deficient pronoun' constraint.

So, why do semantically referentially deficient antecedents favor anaphoric Ø in CVC when Ø subjects are usually associated with highly salient or topic-worthy discourse referents? One reason might be related to a language-specific property of Ø subjects. Another reason could be that this reflex is particular to Lusophone-based vernaculars (given previous findings for Ø in BP). Ariel (1990:108) noted that there should be some language-specific idiosyncrasies in anaphora resolution as far as "[referentially] emptier" antecedents are concerned, but that general implicatures should hold:

> […] we should expect some arbitrary, language-specific decisions, determining how far a language can stretch the use of its various emptier Accessibility Markers. However, language-internally, we should find a reflection of Accessibility Theory, or at least no counterexamples to it. We expect no language to allow poor Accessibility markers with

> less salient antecedents but not with more salient ones. For example, no language should allow third-, but not first- (and second-) person referents to be marked with zeros.

This is the case for Ø subjects in CVC and BP: while they are more probable with 3rd person and/or referentially deficient antecedents, they do still occur with 1st and 2nd person referents. Pending further empirical confirmation, it appears that this property might be a feature of Ø that is generalizable across Lusophone-based vernaculars. Alternatively, this may be a feature of languages in which the same set of overt subject pronoun compete with anaphoric Ø to resume referentially deficient and fully referentially specified antecedents alike.

Jones (2014) points out that even though indefiniteness is encoded morphologically on indefinite determiners (or lack thereof) in indefinite lexical DPs, these referring expressions do not necessarily always introduce new information; indefinite lexical DPs can still reintroduce old referents (second mention). She also shows that they can 'gain' definiteness when occurring in certain discursive contexts and from shared real-world knowledge between interlocutors:

> […] [the indefinite article] is by no means the only factor which contributes to [indefinite expressions'] (in)definite status and felicitous interpretation. The surrounding context and ongoing discourse play a major role in how an expression is interpreted, as well as the intensions of the speaker. (Jones 2014:306).

Therefore, despite some degree of referential deficiency, the discourse context in which an indefinite lexical DP occurs can nonetheless endow it with topical importance or allow it to become highly referentially activated, thus making resumption by anaphoric Ø more probable. Similarly, the distinction between anaphora and cataphora may also suggest why indefinite LEXICAL DPs and similar referring expressions can nevertheless be topical and thus available for resumption by Ø. Givón (2017:3-4,187-188) associates anaphoric Ø with information predictability (150), while cataphoric Ø is associated with topicality and referential importance (151).

(150) **Anaphoric zero and information predictability**

a.) **Communicative:** "Predictable information can be left zero-marked".

b.) **Cognitive:** "Information that is already activated under current focal attention requires no re-activation."

(151) **Cataphoric zero and topic importance**

a.) **Communicative:** "Unimportant information, one that is not expected to persist in the subsequent discourse, can be left zero-marked".

b.) **Cognitive:** "A heightened state of alertness or anticipation needs to be maintained only for the important information that is expected to persist."

While Givón (2017:248) admits that the notion of referential continuity (or accessibility) and referential importance (or topicality) are closely related, he insists that "the two dimensions are distinct and can be dissociated." And of crucial importance for our discussion here, "a referring indefinite NP is by definition maximally discontinuous anaphorically but may be highly topical/important cataphorically or non-topical/unimportant."

The examples of referentially deficient Ø anaphora in the current corpus are not structurally cataphoric (they do not precede their coreferential referring expression), but their antecedents may still bear topical prominence or importance in the discourse. If there is some reduction in accessibility that is incurred from the referential deficiency of inanimate, indefinite, and nonspecific LEXICAL DPs (and similar construction) that makes it less probable for Ø to be realized over an overt form, it may be compensated for by the potential topical importance of a referentially deficient LEXICAL DP or DP + INTERVENING MATERIAL + Ø construction relative to its surrounding discourse. It is also important to acknowledge that this relationship was only active at short anaphoric distances, meaning that any topical importance a referentially deficient LEXICAL DP or DP + INTERVENING MATERIAL + Ø construction may have had tapers off after a few words.

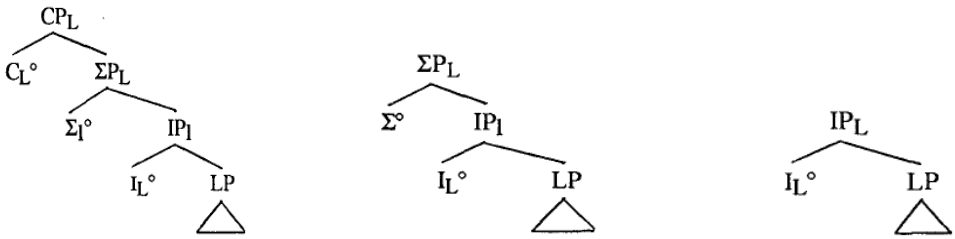
Formal theorizing about *pro* may also provide some answers. The results for specificity and animacy I have found here for CVC reflect what Duarte & Soares da Silva (2016) found for BP. When definiteness is taken into account, these results lend empirical support to that notion that *pro* exhibits the properties of a deficient pronoun, since deficient pronouns are associated with non-human, indefinite, and nonspecific semantic values (Chomsky 1981; Rizzi 1986a; Cardinaletti & Starke 1999:175).

In Section 3, I reviewed the tripartite analysis for CVC subject pronouns in Pratas (2004) and Baptista (2002), which is based on Cardinaletti's & Starke's (1999) Typology of Structural Deficiency. This universalist model asserts that there are three 'abstract underlying classes' of linguistic element[87]: strong, weak, and clitic. The properties of each category are a consequence of the degree to which they are structurally deficient. Strong and weak forms are XPs (they occupy a specifier position), while clitics are X° (in a head position), though weak forms demonstrate a reduced number of the semantic, morphosyntactic, prosodic, and discursive properties relative to strong forms. This 'deficiency' is attributed to an actual lack of underlying structure, weak forms have less functional projections than strong forms, and clitics have less than weak pronouns (152). In Table 21 above, (Section 4.1), it was observed that CVC disyllabic tonic subject pronouns are taken to be strong while their monosyllabic counterparts are taken to be weak[88].

87 The typology of structural deficiency does not apply only to pronouns but is extended to other parts of speech as well.

88 Although the strong-weak distinction was not represented in the coding scheme for X2SBJs in the present study.

(152) a. Strong b. Weak c. Clitic



While a full discussion of the properties of each element is beyond the scope of this chapter (see Section 4.3), the semantic properties associated with each class of pronominal are important. Strong pronouns are more referential, they tend towards animacy, definiteness, and specificity when compared to deficient forms. Deficient forms (weak and clitic) receive their referential properties from antecedents, and given their own structural deficiency, they should tend towards referentially deficiency, expressing semantic values like inanimacy, indefiniteness, and nonspecificity. This was partially borne out by my findings: anaphoric Ø competes with SCs to resume antecedents associated with inanimate, indefinite, and nonspecific antecedents. This effect was strong enough that it overrode the influence of priming (the most important predictor) at short anaphoric distances. The correlations were strongest for animacy and definiteness.

Recall that X2SBJs always employ either a disyllabic tonic pronoun (*ami, ael*, etc.), or a monosyllabic tonic pronoun (*mi, el*, etc.), in addition to a SC, while the use of singleton tonic pronouns (with all verbs except copular *e* 'to be') is residual in CVC (7 out of 3,995 tokens = 0.18%). X2SBJs never occurred with an inanimate antecedent referent, only twice with an indefinite antecedent, but did occur with nonspecific antecedents (n = 36 = 13% of all X2SBJs). In instances of X2SBJs with nonspecific reference, they were almost always used when contrasting and alternating back-and-forth between referents. In (153), the speaker introduces a nonspecific referent indexed with the subject relative *kada algén […] ki teni kantu trabadju* 'each person that has so much work'. This referent is coindexed with two subsequent object clitics *-l*, and then is

shifted to the role of subject in the next clause, in which coreference is born by third person SC *e*. Finally, a coreferential WEAK + SC X2SBJ construction *el=e* is used to establish contrast with the subsequent, disjunctively referential nonspecific DP *kel otu* 'that other person'.

(153) *N=ta pui [kada algén […] ki teni kantu trabadju]$_i$, N=ta*

1.SG.CL=TMA put [each person COMP has so much work]$_i$ 1.SG.CL=TMA

dividi=l$_i$, N=ta po=l$_i$ k-um trabadju.

divide=3.SG.CL.ACC$_i$ 1.SG.CL=TMA put=3.SG.CL.ACC$_i$ with-DET job

e$_i$=ka podi sata fazi kantu trabadju kiii

3.SG.CL=NEG AUX TMA do so much work COMP

el=e$_i$ sata ganha txeu, ki kel-otu sata

3.SG.WK=3.SG.CL TMA earn a lot COMP that other person TMA

ganha poku!

earn little

'I would take each person that works a lot, I would divide them up and give them each a job. One can't do so much of the work [so] that that other [person] earns so little!'

For reasons of statistical testing (and following non-formalist categorizations of CVC subject pronouns, e.g. Veiga 1995; Quint 2000a,b; Lang 2012), X2SBJs with disyllabic tonic pronouns were coded in the same category as X2SBJs with monosyllabic tonic pronouns. Therefore, the question of whether disyllabic (strong) and monosyllabic (weak) forms differ with respect to the semantic referential deficiency of their antecedents is one that awaits empirical testing. For now, it can be asserted that X2SBJs in CVC appear to be associated with the semantic properties of strong pronouns under the Typology of Structural Deficiency: both kinds of X2SBJ never (or barely) admitted inanimate or indefinite antecedents and were mostly used contrastively when bearing nonspecific reference. Since Ø itself is assumed to be referentially deficient (cf. Kato &

Duarte 2003, 2005; Duarte & Soares da Silva 2016), it competes with SCs and is promoted with antecedents that are also referentially deficient (in collusion with other discourse-organizational properties like linking and anaphoric distance). Given that SCs occur productively with a range of antecedent types, probabilistically, Ø can be found to be favored since it has a specific connection to semantic referential deficiency.

It is possible that the CVC system is the eventual outcome for partial Null Subject Languages (NSLs) that continue to transition away from consistent NSLs. The first step in this diachronic path of change involves erosion of person-number inflectional morphology that identifies Ø, and this interacts with reorganizational processes in the subject pronoun system (cf. Duarte 1993, 1995 for BP; Toribio 2000 for Dominican Spanish; Camacho 2008, 2016 for BP and DS). If inflection becomes (nearly) completely eroded, tonic pronouns will grammaticalize into atonic SCs, which themselves can become reduced into bound inflection morphemes (Givón 1976, 1983b, 2017; Hopper & Traugott 1993:7; Harris & Campbell 1995; Siewierska 2004; Wratil 2011). Alternatively, if morphosyntactic restructuring causes person-number inflection to be lost because the base form of the verb is modeled after a nonfinite form in a contact vernacular's source language (as is the case with CVC; Quint 2008b), the same grammaticalization 'pull' on the subject pronoun system will obtain.

Grammaticalization on SCs is catalyzed by a diachronic increase in reliance on topic-shifted constructions with left-dislocated lexical DPs, which slowly transform into 'true doubling structures' (in the sense of Culbertson 2010) as SCs progressively become more affix-like ('the NP-detachment Hypothesis) (Givón 1976) (see Section 4.4). Presumably, X2SBJs (tonic pronoun + SC) also become increasingly possible as SCs become more affixal. This very process has been documented in the history and Gallo-Romance, Rhaeto-Romance, and Gallo-Italic Romance (Roberts 1993; Kaiser 2009; Poletto & Tortora 2016, and references therein). Indeed, Kato & Duarte (2003, 2014b), Duarte & Kato (2014), and Duarte & Soares da Silva (2016:23) have concluded that BP may be developing in the direction of a more CVC-like system:

> V initial sentences with null referential subjects can be considered residual cases of an advanced process of change. […] the weakening inflectional morphology allows weak, clitic-like nominative pronouns adjoined to the head of TP and leave its specifier (Spec, TP) free to receive topic-like subjects.

Note that this view effectively constitutes a tacit endorsement of the 'inflectional affix hypothesis' (in which SCs are inflectional heads) for CVC-like subject pronoun systems. Along the same lines, Luís & Kaiser (2016) have concluded that reduced forms of BP subject pronouns like *você* > *cê* demonstrate evidence of bearing some intermediate status between independent pronouns and SCs. It has also been argued that BP increasingly displays the discourse patterns of topic-prominent languages (Pontes 1987; Negrão & Viotti 2000; Modesto 2008; Oliveira da Silva & Álvez Fonseca 2018). I will return to this discussion of diachronic change, SCs, and the Null Subject Parameter (NSP) at the end of this chapter when reflecting upon what the results in the current study for Ø subjects and X2SBJS can tell us about the cross-linguistically disputed status of SCs.

To summarize the discussion thus far, we have seen that a generalized cognitive constraint like structural priming (related to working memory, implicit learning, and referent saliency/discourse activation) is the most powerful force intervening to condition the variable selection of nominative anaphora in CVC, but that the semantics of referentially deficient antecedents have implications for Ø subjects in CVC (as in BP). Whereas Ø anaphora are usually promoted by highly referentially salient antecedents (they tend to be animate, definite, specific), there is a language-specific constraint in CVC (and BP) that links referential deficiency to Ø subjects. Alternatively, it may be that indefinite referring expression can nonetheless be topically important and that topical importance is modulated by the discourse context. Yet another explanation could be that, in languages for which the same set of overt pronouns compete with Ø to resume referentially fully specified and referentially deficient antecedents alike, Ø will be favored when associated with the referentially deficient semantic values. In any event, the claims for the semantic properties of morphemes characterized by formal structural deficiency under the Typology of Structural Deficiency are supported by the findings in the present study if X2SBJs are

taken to be similar in status to strong pronominals. Further, this approach already considers Ø subjects (*pro*) and SCs to be inherently referentially deficient, hence why they compete to resume antecedents that are themselves referentially deficient.

These conclusions also serve as a unique example of how 'Probabilistic Grammar' approaches (Claes 2017) to morphosyntax can complement (and be complemented by) formal theoretical constructs in the GG or Minimalist traditions, much in the way that was envisioned by Tarallo (1987) and Tarallo & Kato (2007 [1989]). Understanding the full range of implications in CVC associated with predictors like SURFACE FORM OF THE ANTECEDENT, PERSON/NUMBER, ANAPHORIC DISTANCE, and the semantic properties of subject referents like ANIMACY and SPECIFICITY/DEFINITENESS, requires that one account for the probabilistic effects of domain-general cognitive processes like priming, along with well understood linguistic constraints related to antecedent activation and salience (such as accessibility hierarchies). At the same time, one must not reject *a priori* classificational schemes like the Typology of Structural Deficiency and related claims for the properties of the morpheme underlying phonetically unvalued subjects, *pro*, otherwise one is forced to produce language-specific stipulations that diverge from the usual assumptions of models like Accessibility Theory (Ariel 1990).

Of course, in relying upon explanations from GG traditions in which underlying morphosyntactic forms are often taken to be categorical in their properties, there may arise a theoretical contradiction, since a probabilistic or gradient view of the properties of linguistic forms remains a requirement of Probabilistic Grammar. In a sense, this may not be incompatible with the original design of the Typology of Structural Deficiency, as Cardinaletti & Starke (1999:213-214) did leave some room for gradience and probabilistic relations regarding asymmetries between strong, weak, and clitic forms:

> […] while these asymmetries seem to be universal, none of the interpretive asymmetries is systematic: It is not the case that there is a strict covariation between being of one class, and having one type of semantic/phonetic interpretation. The interpretational characteristics are asymmetric but overlapping: the three classes are purely abstract (for example, both deficient and strong elements can refer to human entities and to prominent

> discourse referents, although an asymmetry holds with respect to non-human entities and non-prominent referents. (Cardinaletti & Starke 1999:213).

As far as the NSP is concerned, I have already described how this typological model could benefit from incorporating some mechanism to account for the interfacing between language-internal and language-external processes for phenomena like SPE. This would entail a full account of the interactions between domain-general cognitive constraints (priming, anaphoric distance), constraints involving symbolic conceptual systems (the semantics of subject referents), language-specific constraints like the morphological composition of pronominals (person/number, type of anaphoric expression), and their syntactic role (subject/object/oblique etc., clause type, anaphoric binding), which together determine referential salience and antecedent accessibility.

Clearly, SPE in CVC (and in other languages where Ø anaphora do not require identification from 'rich' inflection) cannot be understood without accounting for all these considerations. At the end of this chapter, I will return to reflect on the descriptive and explanatory soundness of the NSP when discussing what the results from the present study can tell us about the typologically contested status of SC and how SCs fit into cross-linguistic models of SPE.

7.1.3 Inter-clausal relationships, clause type, and turn taking.

Three levels of the constraint ANTECEDENT ACCESSIBILITY PATTERN were significant in MLR 2 (Analysis 2) and strong effects were found for this constraint in the binomial mixed effects model that compared X2SBJS against ALL ELSE (Ø and SCS) (Model 4, Analysis 4). In the binomial mixed effects model that compared Ø against overt forms (Model 3, Analysis 3), this factor was not selected from the model building and selection procedure. In MLR 2, PATTERN G and PATTERN D contexts were found to significantly disfavor Ø subjects. Both switch-reference contexts, PATTERN G was when the antecedent was in a 'discourse chunk' separate from that of the target, and PATTERN D was when the antecedent was the subject of a non-adjacent preceding clause. PATTERN D was the only significant favoring context for X2SBJS in MLR 2. In MLR 4, all ANTECEDENT ACCESSIBILITY PATTERNS in which the antecedent was in a non-adjacent prior clause

(PATTERNS D, E, and G) – also all switch-reference contexts - were found to promote X2SBJs. The only subject-reference-switch context that did not significantly correlate with X2SBJs was PATTERN C, which was when the antecedent was in the immediately prior adjacent clause but had some syntactic role other than that of the subject.

From these results, the generalization can be made that X2SBJs are used mostly to resume referents in non-adjacent clauses; they reintroduce old discourse referents. When X2SBJs occur in proximity to other x2SBJs and referring expressions that are similarly prominent in the discourse like lexical DPs, their discursive function is to mark contrastive focus. This lends further support to the notion that X2SBJs are functionally similar to singleton tonic subject pronouns in languages that rely less (or not at all) on such X2SBJ constructions (i.e. languages in which atonic subject pronominals may be less grammaticalized when compared to SCs in languages like CVC). As has been noted, in languages with tonic and atonic pronominals, and Ø anaphora, the latter two typically compete for selection in many of the same contexts (e.g. discourse chain-medial position, where there is maximal referential continuity) (Givón 2017:283-284). Ø is associated with the highest degree of referential continuity, followed by atonic pronouns (example 149, Section 7.1.2) (Givón 2017:283), and the syntactic role of one of the devices in (149) as a subject or oblique helps determine the referential continuity of the anaphor. Antecedents in the subject position are the most referentially continuous and topic-worthy, followed by direct objects, and then obliques (Givón 2017:284). Thus, it is to be expected that Ø would be less probable in switch-reference contexts, while X2SBJs - CVC's primary way of employing tonic pronouns - would be more probable in the same contexts.

In languages lacking atonic pronominals, for which agreement may have instead become bound to the verb as person-number inflection (i.e. Spanish, Portuguese), Ø subjects tend to emerge in the same places in which SCs and Ø compete in languages that do have atonic pronominals. For Givón (2017:7), who takes Spanish inflectional morphology to be anaphoric (pronominal) agreement, "anaphoric pronouns and zero anaphora have merged into a single [referential continuity] device, subject pronominal agreement […]".

In CVC the burden of referential continuity is born by both atonic SCs and Ø anaphora; since these tend compete for the same subject positions, Ø emerges only when there is a confluence of predictive factors that together work to maximize accessibility or to activate the semantic referential deficiency that is associated with Ø anaphora and their antecedents in Lusophone and Lusophone-origin vernaculars. SCs are far along enough on the grammaticalization cline (cf. examples 49 and 50, Section 4.4) towards bound person-number markers that tonic pronouns seldom occur in isolation; instead, tonic pronouns double with SCs in most situations (with all verbs except *e*). Rather than redundant morphological marking, the function of X2SBJ amalgamations is equivalent to overt tonic pronouns in other languages, indicating that SCs have lost at least some of their pronominal status. Discussion of the SC-to-bound morpheme grammaticalization cline is revisited in the next subsection.

Other constraints that were related to antecedent accessibility and referential continuity were clausal LINKING, CLAUSE CHAIN POSITION, CLAUSE TYPE, TMA FRAME SWITCH, and TURN taking. Clausal LINKING was ranked the second most important predictor in the variable importance plots for Analyses 2 and 3, and was returned as highly significant in the regression models for both. It was not chosen following the model building and selection process for the model in Analysis 4, however.

PROSODIC LINKING was when adjacent clauses/IUs containing antecedent and anaphor were 'linked' by continuing intonation (cf. Chafe 1993, 1994; Du Bois *et al.* 1993; Gumperz & Berenz 1993; Torres-Cacoullos & Travis 2019). In MLR 2, PROSODIC LINKING was a favoring context for Ø subjects and a disfavoring context for X2SBJs, and in binomial Model 3, PROSODIC LINKING was also found to promote anaphoric Ø. SYNTACTIC LINKING was when adjacent clauses containing antecedent and anaphor were linked by a coordinating conjunction, complementizer, or discourse markers (Sections 4.5.3 and 5.5.1). Contexts in which antecedent and anaphor in adjacent clauses/IUs were BOTH prosodically and syntactically linked (by conjunctions and continuing-final intonation), also emerged as a favoring context for Ø (in Model 3). Torres Cacoullos & Travis

(2014, 2015, 2016, 2019) have found comparably strong results for LINKING in English and Spanish.

Unsurprisingly, results for MLR 2 revealed that X2SBJs were disfavored in prosodically linked contexts, although LINKING was not included in Analysis 4 following the model selection process. Crucially, after inspecting interactions among constraints using conditional inference trees (Section 6.2-6.3), it was revealed that LINKING activated the priming effect and the relationship between semantically referentially deficient antecedents and Ø subjects discussed in the previous subsection. This is likely because LINKING enhances referential continuity and antecedent accessibility. Speakers with higher SESSCOREs are stricter in their adherence to the LINKING constraint: they accept Ø subjects in less unlinked contexts than speakers with lower SESSCOREs. Like priming, linking is a good candidate for a constraint that is active for SPE cross-linguistically. It remains to be seen if, in languages for which the discourse carries the majority of the burden of identifying anaphoric Ø (e.g. CVC, English, 'creole' languages, among many others), processes like priming and linking might be even more impactful than in languages for which the burden of identifying Ø is borne by bound person-number inflection.

The factor CLAUSE CHAIN POSITION was included to account for the probabilistic effect on anaphora resolution of the anaphor's position a clause chain or 'discourse chunk' (cf. Section 5.5.2.1). Units of conversational speech tend to be organized under a unifying intonational and thematic contour (Givón 2017:243) and the position of an anaphor in a sequence of chained clauses will be deterministic for SPE. Both SCs and Ø anaphora are high referential continuity devices; Ø subjects are disfavored at clause-chain boundaries (where referential continuity is reduced) relative to atonic pronouns, but are promoted clause-chain medially (where referential continuity is maximal). SCs and tonic pronouns compete at clause-chain boundaries, such as CLAUSE CHAIN INITIAL positions, but stressed pronouns will be selected over SCs to mark switch-reference, establish contrastive focus, reintroduce an old discourse referent, or for stylistic emphasis.

These observations are reflected in the results for CLAUSE CHAIN POSITION in CVC. In MLR 2 and Model 4, CHAIN INITIAL position was a favoring context for X2SBJs over SCs, and over ALL

ELSE (SCS + Ø), respectively. CLAUSE CHAIN POSITION was not one of the predictors chosen from the model selection and comparison procedure in Analysis 3. At the start of a clause chain, referential continuity is low, and old discourse referents are usually reintroduced, thus increasing the likelihood that a X2SBJ will be selected. The results for this predictor support those for LINKING and ANTECEDENT ACCESSIBILITY PATTERN in indicating that, in CVC, X2SBJs are akin singleton tonic pronouns in languages that do not rely extensively on X2SBJ constructions or that do not have atonic pronominals.

The constraint TURN was included to account for the dynamics of the question-and-answer format in the interview task, as opposed to the narrative format of the picture description task. Turn taking has been associated with a lexically-specific effect (for certain verbs) in SPE in Spanish (Bentivoglio 1987; Davidson 1996; Torres Cacoullos & Travis 2014, 2015). In English, the first-subject first-mention position has been found to be a favoring condition for Ø (Wagner 2016), and Torres Cacoullos & Travis (2019) found an IU-initial restriction on Ø in English (in the absence of a coordinating conjunction).

In the question-and-answer format of an interview, clause chains or discourse chunks were bookended by the interviewers' questions and respondents' closures. The RESPONSE level of the predictor variable TURN was assigned to anaphora in the first IU in a speaker's answer. In the PICTURE DESCRIPTION NARRATIVE, each 'discourse chunk' was organized by the speaker (usually corresponding to a full description of the scene on a single page), rather than being prompted by a question. The level INITIAL accounted for anaphora in the first IU of a new, speaker-initiated discourse chunk, separately from RESPONSEs to interviewers' questions. Therefore, most of the anaphora coded as having occurred in INITIAL position were from the PICTURE DESCRIPTION NARRATIVE, but when a speaker initiated a new discourse chunk (unprompted by a question) during the interview task, any anaphora in the first IU of that discourse chunk were also coded as INITIAL.

To reiterate, in the present study, the constraint TURN had three conditions: the level RESPONSE was for targets in the first IU in an answer to an interviewers' question, TURN INITIAL was for an anaphor in the first IU in a discourse chunk that was initiated by the speaker without

being prompted by a question, and TURN INTERNAL contexts were simply when the participant continued speaking after a RESPONSE or TURN INITIAL IU.

In MLR 2, there was a significant disfavoring effect for Ø subjects in TURN INITIAL clauses. TURN was not one of the predictors selected for Model 3, but the RESPONSE context exerted a significant favoring effect on X2SBJs in Model 4. The disfavoring effect on Ø in TURN INITIAL contexts obtained because, at the start of a new 'discourse chunk', discourse referents have been deactivated since the closure of the previous clause chain; the old discourse referents are thus inaccessible (or nearly so), making it difficult for them to engage in a coreferential relationship with anaphoric Ø, thus making the use of an overt anaphor more likely. In question responses, speakers are using X2SBJs to establish contrast with some referent introduced in an interviewers' question or to promote an available discourse referent as a topic.

In MLR 2 and Model 3, I observed significant results for the predictor TMA FRAME SWITCH. Switches in the TMA frame from clause-to-clause have been found disfavor Ø and promote overt pronoun in Spanish (Bayley & Pease-Álvarez 1997; Geeslin & Gudmestad 2011). This is likely because of an associated decrease in referential continuity between antecedent and target clauses. Since in CVC there is no one-to-one correspondence between the TMA markers and the actual temporal, modal, and aspectual interpretation assigned to a VP, and because the TMA frame for bare verbs is contingent on stativity, the TMA marking of the verb phrase and its actual TMA FRAME interpretation were recorded as separate predictors.

This latter constraint, TMA FRAME, was not ultimately retained in any of the models since it ranked low among predictors during the model building and selection procedure and since there were already a large number of parameters under consideration. There were also no clear-cut predictions that could be established for the effects of TMA FRAME on SPE in CVC based on the prior literature. Nevertheless, the predictor TMA FRAME SWITCH relied on the coding of TMA FRAME to indicate when there was a shift (or not) in TMA frame interpretation between an anaphor's clause and its antecedent's clause. When an antecedent was not the subject in its containing clause, this was recorded as a separate category (=X).

In MLR 2, TMA frame SWITCHes (=Y) were and a disfavoring context for Ø, while contexts in which the antecedent played some syntactic role in its containing clause other than that of the subject (=X) were a favoring context for X2SBJs. Model 3 revealed that both TMA frame SWITCHes and X contexts disfavored Ø subjects, with a particularly strong effect in the former condition. TMA FRAME SWITCH was not selected from the model building and comparison procedure for Model 4. If switches in TMA frame are assumed to decrease referential continuity, this would explain why anaphoric Ø is less probable in these contexts.

One more language-internal predictor that emerged as significant in more than one of the analyses was CLAUSE TYPE. Predictions for SPE in CVC as an effect of clausal type were unclear. Costa & Pratas (2013), who deny the possibility of *pro* in root contexts in CVC, identify embedded clauses with a subject anaphor that is c-commanded by an indefinite or WH operator as the only context where Ø is possible. On the other hand, all the examples in Baptsita (2002) of Ø subjects were in main clauses in which the Ø anaphor was coreferenced with its immediate antecedent. Lipski (1999) provided many examples of Ø in main clauses in several Iberian-lexifier creoles but calls into question the degree to which these can be attributed the status of *pro*, since many of these were used with generic or impersonal reference (pro_{ARB}), or at least were ambiguous between generic and genuinely referential readings. Similar observations regarding ambiguously generic vs. referential readings of Ø have been made for Papiamentu (Kouwenberg 1990). However, Kouwenberg observed the subject of a *pa*-clause was a site of variable SPE in Papiamentu, and that Ø subjects in these contexts could be attributed the underlying status of *pro* or pro_{ARB}. Mufwene (1988) offers many examples of Ø in main clause contexts in Gullah, at least some which are unambiguously referential and anaphoric.

The predictor CLAUSE TYPE included levels for ADVERBIAL SUBORDINATES, *PA*-CLAUSES, conditional *SI*-CLAUSES, SUBORDINATE clauses (with complementizers *ki* and *ma*), RELATIVEs, COORDINATE clauses, and QUESTIONs. In MLR 2, the embedded contexts - ADVERBIAL SUBORDINATE, RELATIVEs, and SUBORDINATEs (*ki* and *ma*) - all disfavored Ø subjects. Each of these remained disfavoring contexts for Ø in Model 3, while *pa*-clauses emerged as a favoring

context for anaphoric Ø in Model 3. CLAUSE TYPE was not selected as one of the predictors for the model building and comparison procedure for Model 4.

It is unclear why the only embedded clausal configuration in which Ø is promoted is in *pa*-clauses, while it is disfavored in all other embedded contexts. It could be related to scalar differences in finiteness associated with different kinds of embedding (cf. Nikolaeva 2007; Estrada-Fernández 2016). A closer inspection of Ø with *pa*-clauses, and the differences and similarities in patterns of SPE among *pa*-clauses in Papiamentu and CVC, are topics of interest for further investigation. The general trend for CVC - and for other languages in which the burden of identification of Ø subjects comes from nonlocal antecedents (discourse identification) rather than local person-number inflection markers (or other formal mechanisms like those associated with binding relationships) - appears to be that Ø is favored in main clause contexts.

7.1.4 Summary of the major conclusions: the results for null subject/anaphoric zero and double-subject constructions.

Before moving on to a discussion of what the results for Ø subjects and X2SBJs can tell us about the status of SCs, I take stock of the discussion thus far. The major conclusions that can be drawn are:

(i) Structural priming is the most important constraint in CVC for the realization of both Ø subjects and X2SBJs.
   a. At short anaphoric distances, the Ø-to-Ø priming effect is activated, while X2SBJs are primarily used contrastively.
   b. X2SBJ-to-X2SBJ priming happens at longer distances, across non-adjacent clauses, since X2SBJs serve to reintroduce old discourse referents.

(ii) In CVC (like in BP) Ø subjects are associated with antecedents that are semantically referentially deficient (i.e. inanimate, indefinite, nonspecific antecedents) and 3rd person reference.

a. This relationship is unlocked by prosodic and syntactic linking across adjacent IUs with coreferential antecedent and anaphor.
b. Speakers with lower SESSCOREs are less sensitive to (ii.a).
c. Like in BP, an "avoid referentially deficient pronoun constraint" (Kato & Duarte 2003, 2005) is active in CVC, if this constraint is taken to apply probabilistically.
d. The relationship in (ii.c) might be unique to Lusophone-based vernaculars, or perhaps to languages where the same set of overt pronominals that resume referentially fully specified antecedents, also compete with Ø to resume referentially deficient ones.

(iii) Ø subjects are disfavored in embedded contexts, except *pa*-clauses (like in Papiamentu).

(iv) Ø subjects are disfavored when there are TMA SWITCHes between clauses, presumably because this decreases referential continuity and antecedent accessibility.

(v) X2BJs are favored when their antecedents are in a non-adjacent clause. This is because they are switch-reference devices and are used to introduce old discourse referents.

(vi) X2BJs are associated with the semantic values of strong forms under the Typology of Structural Deficiency (animacy, definiteness) but are used with nonspecific referents contrastively.

(vii) X2BJs are used in responses to questions so they can establish contrast with a competing referent introduced in the question, or to promote the topicality of an antecedent.

(viii) X2SBJs appear to have a similar discursive and functional properties to singleton tonic pronouns in languages that do not rely extensively on X2SBJs.

The next major research question in the present study was concerned with the status of atonic SCs in CVC. Robust results were observed for the other two subject elements in competition with SCs in CVC and these help to draw some preliminary conclusions about the status of SCs, both in CVC and cross-linguistically. One language-external predictor that emerged significant across

Analyses 2, 3, and 4 has yet to be discussed: SESSCORE. This was an individual-specific scoring of the participants' educational and occupational status. This final predictor is considered after the discussion on the status of SCs.

7.2 SUBJECT CLITICS AND TYPOLOGICAL MODELS RELATED TO SUBJECT PRONOUN EXPRESSION/ANAPHORA RESOLUTION.

We have seen that SCs are the default subject type in CVC, appearing across morphosyntactic, discursive/pragmatic, and semantic categories (for all verbs except *e* 'to be'). An application of the tests usually used to distinguish clitics from affixes is beyond the scope of the current study (cf. Kayne 1975; Klavans 1982; Zwicky & Pullum 1983; Rizzi 1986b; Bresnan & Mchombo 1987; Cardinaletti & Starke 1999; Baptsita 2002; Pratas 2004; Siewierska 2004; DeCat 2005; Culbertson 2010; *inter alia*), but needless to say, attempts at drawing such distinctions have not resulted in a consensus as to the status of SCs in CVC (cf. Baptsita 1995, 2002 vs. Pratas 2004), nor cross-linguistically (cf. Poletto 2000; Culbertson 2010; Poletto & Tortora 2016). Thus, some other means of assessing the pronominal or affixal status of SCs might be necessary.

Culbertson (2010:117), following Givón's (1976) 'NP-Detachment Hypothesis', took high rates of X2SBJs in Colloquial French to be indicative of the on-going grammaticalization of SCs as a result of "overuse of stylistic constructions like dislocation." A diachronic increase in the reliance on topic-shifted structures like (example 51, above; recast here as 154)[89] can cause such doubling constructions to become pragmatically neutralized (Siewierska 2004:263).

(154) *Amilson, e ta kumi txeu dimâs*
A. 3.SG.CL TMA eat a lot too much
'A., he eats way too much.'

89 The comma indicates a pause in intonation between the lexical DP and the SC.

The functional/discursive neutralization of these structures catalyzes the grammaticalization process on independent pronouns, pushing them closer to SC status, and eventually affixal status: "such topic-shifted constructions become reanalyzed as neutral clauses; the left- […] detached topic becomes the subject and the anaphoric person marker becomes attached to the verb." (Siewierska 2004:263) (example 52, above; recast here as 155)[90].

(155) *Amilson e=ta* *kumi txeu dimâs*
A. 3.SG.CL=TMA eat a lot too much
'A., he eats way too much.'

Presumably, as SCs proceed further along the grammaticalization cline and are increasingly analyzed as agreement affixes, then tonic pronouns, like lexical DPs, come to participate in X2SBJ constructions as well. As we have seen from the results of the present study, X2SBJs in CVC (tonic pronoun + subject clitic) have become the primary way to deploy a tonic pronoun in this language. This fact itself may be evidence that CVC SCs are grammaticalizing towards bound (affixal) person agreement markers and losing their pronominal status.

This transitional stage in the diachrony of person agreement markers may be the clue as to why it has been so hard to achieve a consensus on the morphosyntactic status of SCs in CVC and cross-linguistically. Part of the confusion comes from theory-specific tests or criteria for distinguishing independent clitic pronominals from bound agreement affixes. Siewierska (2004:121) pointed out that when it comes to distinguishing 'person agreement markers' (~inflectional affixes) and 'anaphoric pronouns' (~phonological clitics), "the distinctions made have been based on theory-internal grounds", and not just in GG approaches, but in functionalist or usage-based approaches too. Under the latter theoretical paradigms, many have adopted the view that affixal person markers (e.g. bound person-number morphology in

[90] The lack of a comma indicates no pause in intonation between the lexical DP and the SC, but instead a unbroken intonational contour.

Portuguese & Spanish), and anaphoric pronouns (e.g. pronominal clitics), are essentially the same 'kind' of person marker, and only differ with respect to their 'domain of agreement' (Siewierska 2004:122). Following the logic of this distinction (see also Lehmann 1982; Bresnan & Mchombo 1987), when person markers are in local agreement configurations where they belong to the same syntactic constituent as their 'controller' (e.g. inflectional affixes engaging in person-number agreement) they are subject to 'grammatical agreement'. When person markers are in non-local agreement configurations where they do not form part of the same syntactic constituent as their controller (e.g. pronominal subject clitics) they are subject to long-distance/nonlocal 'anaphoric agreement' (Siewierska 2004:122-127).

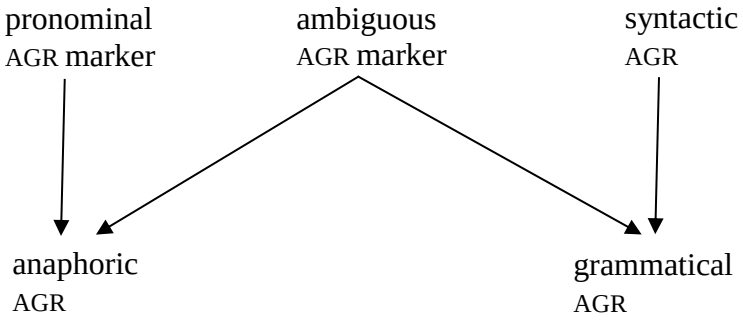
In many languages, there is apparently no distinction between person markers that engage in one or the other type of agreement. To illuminate, Siewierska (2004:122) provides an example from Gumawana (Olson 1992:326, 306); in (example 53, above; recast here as 156a) the person marker *i* might be said to be a grammatical agreement marker given the presence of the lexical DP, whereas in (156b), we appear to be dealing with a anaphoric agreement person marker.

(156) a.) *Kalitoni* ***i****=pasewa*
K. **3.SG**=work
b.) ***I****=situ* *vada* *sinae-na*
3.SG=enter house inside-3SG(INAL)

Of course, this is exactly the predicament we have seen for the debate over SCs in CVC. (156a) is equivalent to a X2SBJ construction in CVC, while (156b) is equivalent to a singleton SC subject in CVC. For Siewierska (2004:126), these kinds of person markers, which seem to lead a 'double life' as pronominal-like anaphors or inflectional-like affixes, are really one and the same element, 'ambiguous agreement markers'. They can engage in anaphoric (non-local) agreement, as well as grammatical (local agreement). Cross-linguistically, ambiguous agreement markers are distinguished from (unambiguous) pronominal agreement markers or syntactic agreement

markers. This makes for a tripartite typological distinction in the types of agreement marker available and a bipartite distinction in the kinds of agreement in which person markers can engage (Figure 56) (Siewierska 2004:126).

Figure 56. "Relationship between types of agreement marker and type of agreement." (Siewierska 2004:126), repeated.



Siewierska (2004:126) also notes that ambiguous person markers are the most common type cross-linguistically. Crucially, "the distinction between pronominal and ambiguous agreement markers and thus between anaphoric and grammatical agreement is a scalar one." (Siewierska 2004:126). Person agreement is scalar because in some languages, it is the properties of the controller that determine when and where agreement markers can occur, and those properties usually align with the accessibility hierarchies of the kind proposed by Givón (1983b, 2001, 2017) and Ariel (1990). Siewierska (2004:149) provides a set of "decomposed" sub-hierarchies that reflect the notions assumed in such accessibility hierarchies (157).

(157) a.) the person hierarchy

$1^{ST} > 2^{ND} > 3^{RD}$

b.) the nominal hierarchy

pronoun > noun

c.) the animacy hierarchy

human > animate > inanimate > abstract

d.) the referential hierarchy

definite > indefinite specific > nonspecific

e.) the focus hierarchy

not in focus > in focus

Siewierska (2004:149) describes this hierarchy of properties of the controller and its relationship to person agreement markers as follows:

> All the hierarchies define a preference for person agreement when the controller exhibits the characteristics on the left of > as compared to those on the right of >. Thus the expectation is that if person agreement is not obligatory in a language, it will occur with controllers displaying the characteristics on the left-hand side of the hierarchies rather than with controllers manifesting the characteristic on the right-hand side.

The results from the present study showed that, in CVC, SCs were more likely to occur in X2SBJ constructions with the properties on the left-hand side of the hierarchies (i.e. when the controller, a tonic subject pronoun, appeared locally), and are more likely to be dropped (Ø), when the controller was a nonlocal antecedent (i.e. a discourse antecedent in a prior IU/independent clause) and bore one of the properties on the right-hand side of >[91]. This is further evidence for the notion that CVC SCs are ambiguous person markers developing along a grammaticalization cline somewhere between independent anaphoric pronouns and bound person agreement affixes. It also provides a rationale for why Ø would be promoted with referentially deficient antecedents: if SCs are in the process of becoming bound person agreement markers, then they are probabilistically

[91] The only modification would be the (157d), nonspecific and indefinite specific would have to change orders according to the results found in the present study.

dropped depending of the deficient referential properties of their controller. This explanation is typologically consistent while also providing a reason for why ascertaining the underlying morphosyntactic category for SCs has proven elusive.

Still, there are some indicators that CVC SCs may be at the early stages of the pronoun-to-clitic-to-affix grammaticalization cline associated with ambiguous person agreement markers. In diachronic perspective, we have noted that this grammaticalization cline is a cross-linguistically common one in the history of many languages (cf. Hopper & Traugott 1993:6-7). This cline can be characterized by changes in formal categories as tonic pronominals become atonic clitics, bound inflection, and eventually Ø. But grammaticalization clines also involve changes in the functional (a) and semantic (b) status of the affected items (Siewierska 2004:261-262): (a) a change in function from "a referential expression with deictic or anaphoric force to a syntactic agreement marker" (158); and, (b) a loss of semantic value, information about the "referential identity of the person markers." (159).

(158) pronominal agreement marker > ambiguous agreement marker > syntactic agreement marker

(159) many semantic distinctions > fewer semantic distinctions

1st vs 2nd vs 3rd	1st & 2nd vs 3rd
Inclusive/exclusive	no distinction
Singular/plural	no number
masculine/feminine	no gender
honorific markers	no honorific markers

Of upmost importance for the discussion at hand on SCs in CVC is that:

> […] the three types of changes do not always coincide […] both clitics and affixes may function as pronominal and ambiguous agreement markers. And in languages with both

> clitics and affixes, it is not always the case that the clitics express more semantic distinctions than the affixes. (Siewierska 2004:262).

I posit that the seemingly contradictory behavior of SCs in CVC, and clitics more generally, is due to their very intermediate status on this three-way grammaticalization cline, along which they occupy a space between formal, functional, and semantic categories. In CVC specifically, SCs have begun to experience grammaticalization as pronominal anaphoric agreement markers but appear to have stagnated at an early phase of the cline. They retain many features of fully referential pronominals, for example, they retain all the semantic values in (159) except gender[92]. On the other hand, like an ambiguous agreement marker, SCs repeatedly emerge in X2SBJs as a seemingly redundantly-marked nominative argument of the verb. These amalgamated forms arise in precisely the same contexts associated in which overt tonic subject pronouns tend to occur cross-linguistically (switch-reference, contrastive focus, stylistic emphasis, animate and definite reference).

SCs are dropped only when the processing burden associated with antecedent accessibility and identification is minimal (as a probabilistic effect of a confluence of factors), or in contexts in which they have become associated with a referentially deficient antecedent (controller), either because CVC is of the typological class of languages where ambiguous person markers are optional based on the semantic properties of the controller, or because of the referential properties associated with the deficient pronoun *pro* under the Typology of Structural Deficiency. Further evidence for the partially grammaticalized status of SCs is of the distributional sort that has already been repeatedly observed (Baptista 1995, 2002; Pratas 2004) - they attach to various hosts (clitic, not affix-like), but can only be separated from the verb by VP-internal heads (affix, not clitic-like), etc. - but despite the descriptive and empirical accuracy of these tests, they have proven unable to provide a consensus distinction between phonological clitics and bound inflectional affix.

This study has provided evidence from the probabilistic distribution of the forms that 'compete' with SCs in SPE in CVC. Along with evidence from observations drawn from

92 Gender is only marked on honorific/polite second person SCs *nhu* (m) and *nha* (f).

typological surveys of the cross-linguistic properties of SCs (cf. Siewierska 2004), these findings suggest that SCs in CVC are ambiguous person agreement markers: not quite independent pronoun, not quite bound inflection affix, but somewhere in between, and thus capable of engaging in either local or non-local agreement with their controllers. Having drawn some preliminary conclusions about the three nominative anaphora of interest in the present study, I turn now to a discussion of the language-external predictor that emerged as significant across models and analyses.

7.3 PRELIMINARY CONCLUSIONS REGARDING THE EFFECT FOR SPEAKERS' SOCIOECONOMIC STATUS SCORE.

By far the strongest effects from a language-external predictor were from the factor SESSCORE. This predictor was based on the metric in Otheguy & Zentella (2012) and assigned each participant points based on their level of education and their occupation. Scores ranged from two to eight, with eight being the highest, assigned mostly to graduate students and public functionaries. Participants with a score of two had not completed high school and mostly worked odd-jobs, such as day-labor agriculture or washing cars in a city plaza (Table 36, Section 5.6). SESSCORE was the only language-external predictor to emerge significant across models in Analyses 2, 3, and 4. It demonstrated strong effects: speakers with a higher SESSCORE consistently used less Ø and X2SBJs, but more SCs; in other words, they used a reduced variety of subject forms overall. These results are robust, but they should be taken as preliminary evidence awaiting further empirical confirmation before being assumed to be generalizable, for several reasons:

One is the small population sample size of the current study. While responses from 33 participants have clearly revealed robust results, particularly with respect to language-internal factors, an increase of 20-30 participants would make for a population sample in which the researcher could have more confidence with respect to language-external predictors. Further, the per-participant token count varied widely from speaker to speaker (min = 25, max = 230, mean = 110.5), and not all participants in the current study completed the picture description narrative,

thus a greater sample size is need to overcome skewing of the results that might come from such imbalances in the dataset.

Despite these shortcomings, recall that Models 3 and 4 were binomial mixed-effects models with PARTICIPANT included as a random factor. This accounted for the variance associated with idiosyncratic patterns of SPE. Several participants were found to differ from the population-norm in their patterns of SPE for Model 4 in Analysis 4 (patterns of X2SBJs). Yet, SESSCORE remained a significant predictor even after accounting for individual-specific variance, meaning that these results should be taken as a preliminary, but robust indication that higher SESSCORE may be correlated with using more SCs at the expense of other subject pronominal forms in CVC.

Why would this be this case? One reason may be that speakers with higher SESSCOREs are likely to experience greater exposure to Portuguese and use Portuguese more often in their daily lives. This was certainly the case for the graduate students in the study who were enrolled in a program that recruits students from across Lusophone Africa, many of whom are not CVC speakers, and thus the intensive course and lab work was carried out mostly in Portuguese. The undergraduate students are also likely to have more exposure to Portuguese and engage in more use of Portuguese on a daily basis than most non-university educated speakers of CVC, though less so than the graduate students. Similarly, participants like the bank teller from Maio or the retired teacher who had lived and studied in Portugal (both of whom had higher SESSCOREs) were also among the participants likely to have had the most regular exposure to and use of Portuguese.

Recall that Cabo-Verdean Portuguese is typologically and stylistically very close to European Portuguese, and though it has not been explored empirically, Cabo-Verdean Portuguese is probably similar to European Portuguese (EP) in its status as a consistent NSL. It has the full person-number suffixal inflectional paradigm associated with EP and probably many of the associated morphosyntactic reflexes too. It is possible that speakers of CVC who engage in regular, prolonged, and intensive use of Portuguese are more closely associating the functional role and formal status of inflectional morphology in Portuguese with the functional role and formal status of SCs in CVC. In other words, it could be that since these speakers are used to a system of SPE

like that of Portuguese - in which there are no atonic subject pronominals and in which bound person-number inflection capable of identifying Ø subjects is obligatory with finite verbs - they may be more inclined to adopt a inflectional-affix-like analysis of CVC SCs. SCs might just be more inflection-like in their CVC 'grammar'.

The conditional inferences trees constructed in Analyses 3 and 4 also revealed that speakers of different SESSCOREs diverged in patterns of SPE in specific contexts. Speakers with lower SESSCOREs were more inclined to use Ø in the 1st person, or across unlinked or merely syntactically linked clauses, and they were also more sensitive to prosodic linking in the 3rd person when there was no priming effect or effect from referentially deficient lexical DP antecedents. In Analysis 4, speakers with lower SESSCOREs were more likely to use X2SBJs in response contexts even when the antecedent was contained in the adjacent prior clause (in the interviewers' question) and they were more sensitive to the X2SBJ-to-X2SBJ priming effect across non-adjacent clauses. This close inspection of the contexts in which SESSCORE becomes relevant indicate that all speakers rely on the same active constraints to resolve SPE in CVC, but that those with higher SESSCOREs have stricter requirements on where non-clitic subjects can occur: all of the properties found to disfavor Ø and X2SBJs have an increased disfavoring effect on the patterns of SPE in the CVC of speakers with higher SESSCOREs.

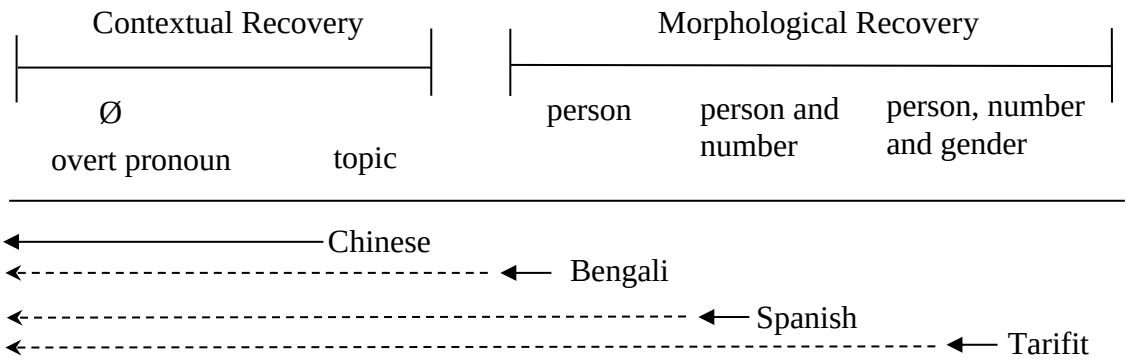
7.4 A REFLECTION ON FORMAL TYPOLOGICAL MODELS OF SPE.

How can the results thus far be reconciled with formal accounts of SPE cross-linguistically? On the one hand, it seems that in light of the typological facts, the NSP is impossible to sustain without some way to account for intermediate, grammaticalized status of SCs, and other properties that resist or have yet to receive formalization, such as antecedent accessibility, structural priming, clause linking, and anaphoric distance. On the other hand, the semantic properties attributed to *pro* as a deficient pronominal, and the categories outlined in the Typology of Structural Deficiency (Cardinaletti & Starke 1999), are at least partially borne out by my

findings, as well as by those of Duarte & Soares da Silva (2016) for BP. Microparametric accounts have been able to capture some of the gradient instantiations of the NSP across canonical and partial NSLs in Romance (Duarte & Soares da Silva 2016), and it would only be a matter of extending this gradient conception of the NSP to languages that have lost bound inflection but have gained atonic subject anaphora.

There have also been robust accounts of Ø subjects in topic-comment or discourse-oriented languages (Huang 1984; Cole 2009; Camacho 2013:146-173; *inter alia*) and for partial NSLs in which the realization of Ø is restricted to particular person-number combinations or certain clausal configurations (Vainikka & Levy 1999; Modesto 2008; Holmberg *et al.* 2009; Camacho & Elías-Ulloa 2010; Wratil 2011; Costa & Pratas 2013; *inter alia*). For Cole (2009:569) and Camacho (2013:112), languages place different "minimal morphological thresholds" on the identification of Ø subjects. Identification proceeds by a "cascading strategy": "[…] null subjects are identified if uniquely identified by agreement morphology. If that fails, they are identified by reference to an antecedent context, and if that is not possible, overt pronouns are used." Languages lie on a gradient scale that places different limitations on the threshold requirements necessary for Ø subject identification (160).

(160) Recoverability scale (Camacho 2013:113)



With respect to the results for the current study, on this scale, CVC would allow for recoverability of Ø with discourse topics like Chinese. If SCs are understood to be ambiguous

person markers, then CVC is like Spanish on the 'recoverability scale' when SCs are engaging in nonlocal 'anaphoric agreement' (recoverability looks to SCs for identification first, if they are not available [for reasons stemming from the confluence of predictive factors in discussed in Sections 6.2 and 6.3.1] then Ø will be identified by a topic). Even in that case, though, the true distribution of Ø in CVC can only be accounted for by considering a combination of probabilistic constraints, as well as the association between Ø and referentially deficient lexical DP antecedents.

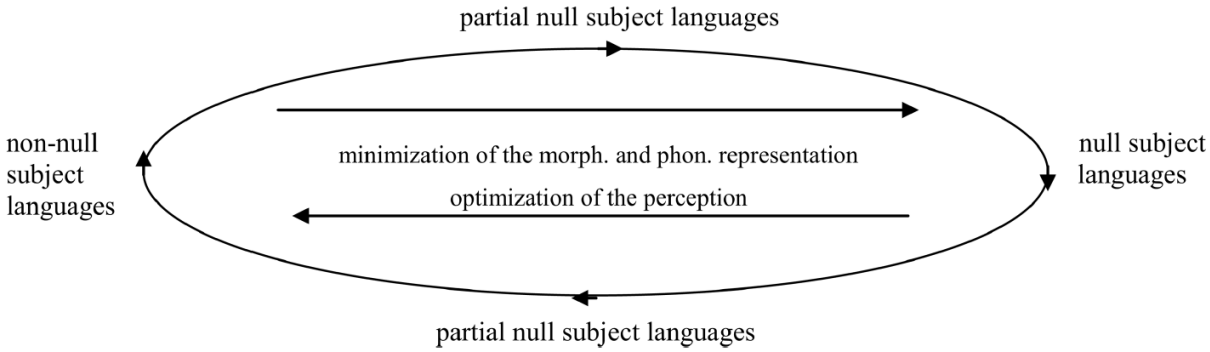
Other recent attempts to integrate quantitative measures of inter- and intra-dialectal morphosyntactic variation with the machinery of the Minimalist Program may also find success in integrating considerations of variable rule systems with assumptions about formal underlying structure (cf. Wilson & Henry 1998; Adger & Smith 2003, 2005; Cornips & Corrigan 2005a,b; Kallen 2005; Sessarego 2011b, 2012a, 2013b, 2014d, 2017a, 2019a; Sessarego & Gutiérrez-Rexach 2010, 2011a,b, 2017, 2018; Gutiérrez-Rexach & Sessarego 2011b, 2014a,b; Sessarego & Rodríguez-Riccelli 2018). Similarly, linguistic-cognitive models based on a modular architecture (c.f. Jackendoff 2015) could provide accounts of how the interfaces between language-specific processes (related to syntactic structure, formal morphosyntactic categories, and the referential semantics of anaphora) and domain-general cognitive processes (priming, anaphoric distance) conspire to determine the effects of constraints like antecedent accessibility.

At least one approach has tried to reconcile the diachronic facts surrounding SCs' erosion into inflectional affixes with the terminology of the NSP (Wratil 2009); although the same author, in later consideration, arrives upon a more pessimistic conclusion regarding the validity of the NSP (Wratil 2011). Wratil examined the interactions between tonic, Ø, and clitic anaphora in the subject systems of various languages. Like Siewierska, she also notes that a large number of languages have atonic subject pronominals, and that many of these exhibit inconsistent patterns of use of Ø anaphora. For example, Ø often replaces SCs (as a results of grammaticalization and eventual complete erosion) in the pronominal paradigm only in certain person-number combinations. Given this and the other typological facts that we have discussed for NSLs, under a

strict, classical formulation of the NSP, most of the world's languages would have to be classified as partial NSLs (Vainikka & Levy 1999).

According to Wratil (2011), the reason for this apparent cross-linguistic inconsistency is due to the fact that speakers, in an attempt to economize the processing burdens associated with tracking referents across discourse, are experiencing a constant push-and-pull between forces that seek to maximize the semantic transparency and discourse-pragmatic salience of subject referents, while also minimizing their morphophonological representation (Figure 3, repeated here as Figure 57). These forces cause tonic subject pronominals to weaken into atonic/clitic pronominals or affixes (ambiguous person agreement markers), and to eventually erode to Ø anaphora, at which point new tonic forms may be introduced into the system (usually they are recruited from other referring expressions Heine & Song 2011).

Figure 57. "The Null Subject Cycle" (Wratil 2011:102), repeated.



This model captures the reality of the grammaticalized nature of SCs and how they might impact the broader subject domain as they erode from independent pronominal to bound affix (and sometimes eventually becoming degraded to Ø). Further, it captures why new tonic pronouns might be [re]introduced into a subject pronoun inventory (by recruitment), and why as SCs erode, they can begin to amalgamate with their tonic counterparts in X2SBJ constructions. The Null Subject Cycle also accounts for the typological reality that most languages are partial NSLs

(cf. Vainikka & Levy 1999, Wratil 2011) and that ambiguous agreement markers are common cross-linguistically (Siewierska 2004). Lastly, it leaves by the wayside the notion that inflectional morphology is required to identify Ø subjects (topics can identify Ø via antecedent accessibility) and finds alternative causal mechanism for explaining why languages change in the NSP status over time.

If this model is on the right track, we can begin to formulate a loose sketch of what the diachronic development of the CVC subject domain might have looked like. In Chapter 3, much of the work of historical reconstruction of the early CVC subject domain that was carried about by Quint (2000a) and Lang (2012) was reviewed; they postulate that a bipartite tonic-atonic subject pronominal inventory emerged early-on in the development of proto-CVC. In Section 3.1, it was assumed that 1st person atonic SCs came about by cross-linguistic structural convergence between morphophonologically similar forms in the superstrate and primary substrate languages; the base form of Mandinka (*ń* > *n*) and Wolof (*nu* > *nu*) person markers would have converged with reduced forms of the tonic pronominals (it is assumed they became reduced due to grammaticalization early-on during the rapid morphosyntactic restructuring that occurred shortly after the settlement of Santiago).

Since the CVC bare verb was based on the superstrate infinitive (Quint 2008b), the person-number inflection of the superstrate would have been eliminated almost immediately during the morphosyntactic restructuring process. It might then be assumed that SCs began to grammaticalize rather quickly, entering the early stages of the grammaticalization cline in which they began to demonstrate some of the properties of an ambiguous person agreement marker. The emergence of the disjunctive tonic-atonic pronoun inventory would plausibly have been facilitated by the presence of such subject pronominal systems in CVC's substrate languages, perhaps as a result contact-induced cross-linguistic structural convergence[93].

[93] Since several Romance vernaculars also have bipartite tonic-atonic subject pronoun inventories, and since such stressed-based disjunctions also existed in the superstrates oblique pronominal system, it might not require much 'restructuring' for such a change to have emerged from intensive contact between languages that reinforce the development of such a system by themselves organizing pronouns along similar lines.

I posit that SCs have likely stagnated at the early phases of the ambiguous person agreement marker stage of the grammaticalization cline. This has resulted in a system in which SCs have become sufficiently grammaticalized that they can amalgamate with tonic pronouns in the majority of cases where tonic pronouns are deployed, and thus X2SBJs have equivalent functional status to overt tonic pronouns cross-linguistically (switch reference, topicalization and focus, stylistic emphasis, reintroducing old discourse referents). Ø subjects have come to respond to two primary functional conditions: they either emerge when the processing burden associated with antecedent accessibility is minimal due to a coalescence of predictive factors mostly related to interactions between antecedent accessibility and domain-general cognitive constraints, or they are sensitive to the semantic referential properties of their controller, or to the formal properties of their underlying morpheme, should this be taken to be *pro* (referential deficiency).

7.5 FUTURE DIRECTIONS.

The present study has established several lines of inquiry that might be pursued in order to better understand the nature of SPE in CVC and cross-linguistically. To begin with, future studies should widen the scope of the anaphora under examination. Tonic, atonic, and Ø anaphora also compete in object position and the scope of analysis should be expanded to account for other clausal arguments. Gathering from casual observation while coding for subject anaphora, object-drop, or anaphoric Ø as an accusative or dative argument, appears to be even more productive than anaphoric Ø subjects.

There is a strong theoretical and statistical motivation for including at least three levels representing outcomes of the response variable in order to conduct a quantitative analysis of SPE in CVC. In the present study, I compared the results from an analysis based on a trinary dependent variable (Analysis 2) with the results from two analyses built upon binary dependent variables (Analyses 3 and 4) in which some levels of the response variable had been collapsed. While this allowed me to pinpoint the factors at play for the non-clitic forms, it also reduces power and

increases the chances of committing Type 1 or Type 2 errors (Pituch & Stevens 2016). Further, the binomial models do not capture the full scope of the variable system with all of its distinctive outcomes. In order to account for the individual-specific and residual variance as is done in mixed-models, the data should be analyzed using a mixed-effects multinomial regression. This will allow for a more parsimonious model of the full variable system underlying SPE in CVC while also according for the variance associated idiosyncratic patterns of SPE. Similarly, the predictor outcomes might be expanded so that there is a category for subject arguments that are singleton tonic pronouns (if their token count is found to increase after expanding the size of the corpus).

Binary variation between Ø and tonic forms with copular *e* 'to be' also needs to be considered from a quantitative perspective, particularly in light of arguments that have analyzed this predicate as the same morpheme as homophonous (except in word-level stress) 3rd person SC *-e-* (Baptista 2004). The challenge in carrying out such an examination would be in distinguishing unambiguously argumental contexts from those with a generic subject or presentational value. Nevertheless, variation between 3rd person, 2nd person, and 1st person plural overt pronouns with generic reference, and generic/impersonal Ø, is also a topic of interest, particularly with respect to the semantics of generic and impersonal reference cross-linguistically.

Closer inspection of the distributional, semantic, and prosodic claims regarding strong, weak, and clitic forms under the Typology of Structural Deficiency may be explored by more carefully examining differences between the disyllabic and monosyllabic tonic forms to see if they correlate with the postulated differences between strong and weak classes. One might also compare the semantic properties of the subject referent across double-subject constructions that are comprised of amalgamated lexical DPs with SCs, particularly with respect to animacy, definiteness, and specificity, similarly to how Tavares Silva, Carvalho, & Ziober's (2018) documented patterns of subject doubling across varieties of Portuguese. Another related direction could be to assess the degree of left-dislocation that can be attributed to tonic subject pronouns engaged in X2SBJs. This could be achieved by comparing the prosodic contour of tonic pronouns in X2SBJs with those of unambiguously left-dislocated lexical DPs. This would serve as a proxy for determining the degree

to which the tonic pronouns in X2SBJs are integrated into the clause as far as their nominative argumental status is concerned (Culbertson 2010). The semantics of the lexical DPs involved in the DP + INTERVENING MATERIAL + Ø construction might be examined to see when, where, and why SCs are dropped in these contexts.

There are many different ways that properties like 'referential coherence' and 'antecedent accessibility' might be interpreted and represented in a quantitative study. Studies like Foley & Valin (1984) and Suckow & Holler (2016) discuss hierarchies of 'coherence' or 'accessibility' that comprehensively combine notions of the semantics of subordination, argument structure, and the sequential and thematic organization of discourse events. These hierarchies might motivate a recoding or merging of various predictors in the inferential analyses. This might be done, for instance, by collapsing into fewer variables the semantic predictors for ANIMACY and SPECIFICITY/DEFINITENESS with ones like ANTECEDENT ACCESSIBILITY PATTERN or SURFACE FORM OF THE ANTECEDENT that are related to priming, the morphological composition of the antecedent, and clausal relations, and these could also be collapsed with discourse-organizational constraints like LINKING and CLAUSE CHAINING.

In order to identify further evidence that CVC SCs are indeed eroding along a grammaticalization cline from anaphoric person marker to ambiguous person agreement marker, a future study might try to identify changes in apparent time in the probabilistic distribution of the various elements involved in SPE in CVC. Changes in SCs that indicate a loss in pronominal status over time, such as in their associated semantic values, distribution relative to other syntactic elements, or erosion in their prosodic integrity, might constitute evidence of a grammaticalization process. Similarly, one might try to identify a diachronic change in the status of SCs by some other means, such as by comparison with other varieties in the Upper Guinea group of Portuguese-lexifier Creoles, some of which are much younger than the Santiago and Maio varieties, like the variety spoken on Santo Antão, and others which have been geographically isolated from CVC for centuries like Papiamentu. By building on Quint's (2000a) and Lang's (2012) historical reconstructions of the early CVC pronoun inventory, one might try to identify the stage at which

new tonic pronouns entered into the system, if this is what triggered the erosion of SCs, and if the rise of the disjunctive subject pronoun inventory can be attributed to substrate influence from the Wolof and Mandinka systems that resulted in cross-linguistic convergence.

In order to conduct cross-linguistic comparisons of subject pronoun systems in which tonic pronouns coexist with SCs, recent proposals for comparative research programs should be used as a guiding framework. These include the 'Variationist Typology' method (Torres Cacoullos & Travis 2019) or the fusion of variationist sociolinguistic and corpus methods with the theoretical underpinnings of the usage-based/functionalist/'cognitive linguistic'/typological tradition into a 'Probabilistic Grammar' research program (Claes 2017). These would be ideal platforms for empirically testing the validity of the Null Subject Cycle (Wratil 2011).

Consideration of extra-linguistic individual-specific predictors might be expanded to better capture the sociocultural categories of particular relevance in Cabo Verde. A more fine-grained representation of Cabo-Verdeans' lived experience might consider time spent abroad or if they have relatives in the diaspora. Additionally, steps might be taken to disentangle issues of socioeconomic status from frequency of exposure to and use of Portuguese. With respect to data collection procedures, it may be useful to replace the 'Frog Story' from the picture description narrative task with a more culturally relevant story for Cabo Verde. For instance, the tale of *Lobu i Xibinhu* 'the wolf and the sheep' will be familiar to almost all CVC speakers and may elicit a more enthusiastic narrative. Additionally, the question battery during the sociolinguistic interview might be expanded to elicit more narrative-oriented story-telling type discourse, which might allow one to tease-apart issues of narrative style and task effects from those concerning person-number or the semantic values of the referents involved in a narrative, as opposed to a question-and-answer or conversational format.

Finally, other methods of ascertaining the mental representation or perception of subject forms in CVC might be considered. With the widespread and growing popularity of social media in Cabo Verde and throughout the diaspora, there is an emergent popular orthography that appears to be rich with variation; it indexes a wide variety of identity-related stances and can provide

unique views as to how native speakers of CVC view parts of speech and the [in]dependence of morphosyntactic constituents. Similarly, laboratory-based means of accessing speakers' grammaticality acceptability judgements may also be used in tandem with less controlled speech data to draw conclusions about the underlying mental representation for anaphoric elements in CVC.

7.7 CLOSING REMARKS.

The present study has offered insights into the nature of SPE in CVC. This study is the first to define the envelope of variation for SPE in CVC In doing so, I have provided a framework that is reproducible and that can be used to test the validity of the delimitation of the envelope of variation and the findings of the analysis. This study is also the first to analyze SPE in CVC using a combination of 'Probabilistic Grammar' (Claes 2017) methodology and theoretical constructions from GG paradigms. Consequently, I have shown that the trans-paradigmatic approach is the best way forward in the quest to arrive upon the most descriptively adequate, theoretically robust, and typologically accurate account of SPE and anaphora resolution.

It was shown that Ø subjects are modulated by many of the same constraints that condition Ø anaphora cross-linguistically, and particularly in languages with little to no person-number subject-verb agreement inflection. These constraints are either directly related to domain-general cognitive processes, the semantics of antecedents, language-specific structural concerns, or some combination of these. Like in BP, CVC Ø subjects are promoted by referentially deficient antecedents. This may be a reflex specific to Lusophone-origin varieties, it may a property of Ø in languages in which SCs are ambiguous person agreement markers that are sensitive to the referential properties of their controllers, it may be related to the properties of the deficient morpheme underlying Ø subjects *pro*, or it may be a reflex of languages where overt pronouns and Ø compete to resume referentially deficient and referentially fully specified antecedents alike. X2SBJ constructions, in which a tonic subject pronoun is amalgamated with a SC as the subject

argument of the verb, perform the same function that overt pronouns do cross-linguistically: they are switch-reference devices, they reintroduce, focalize, and reinforce the topic-hood of some previous discourse referent, or signal stylistic emphasis. Pending further empirical confirmation, it was claimed that SCs in CVC are ambiguous person agreement markers (Siewierska 2004), and that they have become stagnated at the early stages of a cross-linguistically common grammaticalization cline responsible for transforming independent anaphoric pronouns into bound inflectional affixes. It was suggested that the intermediate or in-between status of SCs - not quite pronoun, not quite affix - was the culprit hindering a theoretical consensus on the formal and functional status of SCs both in CVC and cross-linguistically.

Any conclusions regarding the results for language-external predictors are to be taken with caution and less confidence than one might extend to the results for the language-internal constraints in this study. Despite this, robust results were obtained, and these indicate that CVC speakers of higher socioeconomic status use a reduced variety of subject forms, more SCs at the expense of Ø anaphora and X2SBJs. One explanation for this may be that speakers with higher SESSCORES, who are almost certain to use and hear Portuguese on a more frequent and intensive basis, have come to equate CVC SCs more closely with person-number agreement affixes, since the inflectional paradigm is the most frequent mechanism facilitating referential continuity in Cabo Verdean Portuguese.

Finally, this study is situated at the intersection of methodological and theoretical paradigms that too often have been assumed to be incompatible or contradictory. I have drawn on contributions from the GG tradition to construct some of the predictive constraints for a quantitative/variationist sociolinguistic analysis. I also relied on theoretical constructs from other paradigms, including 'Probabilistic Grammar' approaches (Claes 2017). Conclusions drawn from the analysis were used to reflect on theoretical models in all these approaches. This trans-paradigmatic perspective will contribute to a reconciliation between these lines of inquiry and inspire future research that draws on the most successful methodological, descriptive, and explanatory contributions from each of the associated research programs.

Appendix 1: Annotated coding procedure example with Frog Story excerpt

This sample is an excerpt of a transcript from the Picture Description Narrative task that details the coding procedure.

P16 - Male - 19 years old – COMBINED DIALECT REGION (*fora + Santiagu sul*) – SESSCORE = 6 (EDUCATION = undergraduate)

iii keli, keli ta parse=n un joven felis.
Un rapazinhu,
un kriansa, na brinka na bera di agua di-, na bera-, na-, juntu di un arvor ..
juntu di un jardim ..
ki e kuza ki [unnn mininu di CV]$_i$ gosta muitu.
Té albes ***Ø***$_i$ *ta sai sedu.*

1. SPE/anaphoric expression: Ø
   Surface form of antecedent: LEXICAL DP

   Linking: NO LINK
   Person-number: 3SG
   Antecedent accessibility pattern: PATTERN B
   Anaphoric distance: 4
   Animacy: ANIMATE
   Specificity/definiteness: NONSPECIFIC/INDEFINITE
   Turn: TURN INTERNAL
   Clause-chain position: CHAIN FINAL
   Clause type: MAIN
   TMA: *TA*
   TMA frame: NONPAST
   TMA frame switch: NO SWITCH
   Verb class: CLASS 4

Ten otus, N=ten-, [un primu ki ta sai sedu]$_i$,
Ø$_i$ *ta brinka na beira di arvor,*

2. SPE/anaphoric expression: Ø
   Surface form of antecedent: LEXICAL DP

   Linking: PROSODIC
   Person-number: 3SG
   Antecedent accessibility pattern: PATTERN C
   Anaphoric distance: 0

Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/INDEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: X
Verb class: CLASS 4

***Ø**$_i$ ta brinka ku papa lama, lamas di barru asi,*

3. SPE/anaphoric expression: *Ø*

Surface form of antecedent: ZERO

Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 6
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/INDEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

***Ø**$_i$ brinka ku pedra, kaneka, panha objetus, brinkedus, ¿¿??, bonekas, baldi,*

4. This context is excluded from the envelope of variation since the CLASS 4 verb *brinka* is unmarked for TMA (compare with the previous utterance) but cannot be attributed a default past reading.

*i **Ø**$_i$ ba brinka na meiu di rua e otu nem e$_i$ ka ta liga kumida azbes.*

5. This context is excluded from the envelope of variation since its prior antecedent was not an admissible token.

ora di almusu ta sta, ael$_i$, na rua,
*nem banhu **Ø**$_i$ ka ta toma pamo!-*

6. SPE/anaphoric expression: *Ø*

Surface form of antecedent: TONIC
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 4
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE (in 5 the anaphoric chain, which began with the [+specific, -definite] referent *un primu ki ta sai sedu*, switches to

[+specific, +definite] value when the referent is resumed by *e* and then *ael* between 5 and 6.)
Turn: TURN INTERNAL
Clause-chain position: CHAIN FINAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

--- THIRD PARTY PASSERBY BUMPS INTO THE MICROPHONE ---
…
TIMESTAMP: 32:22
…
Ø$_i$ *ta sta na rua na brinka.*

7. Token was not admitted for analysis because of the interruption from the third party (it was treated as following a false start).

Kela mais ou menus e, realidadi di uunnnn, di un kriansa di CV,
Ø *sta na rua na brinka albes,*

8. Token was not admitted for analysis because it appears ambiguously or quasi-finite.

ku animals asi, k-um kãu, ku si brinkedus ou ku otu kolegas na brinka na bera di arvori, asin.

9. The clause chain ends above and a new one begins with the next, clause chain initial Intonational Unit.

Kel$_i$ *e mais ou menus kuza di un, joven kabuberdi-, mininu kabuberdianu,*

1. The above clause is a 'grounding clause' with respect to CLAUSE CHAINING, but it does not contain a potential token anaphor.

mais ou menus ***Ø*** *e keli.*
Ø *e brinka, na rrrua asi, ku kaneka,*

2. Subject arguments of the verb *e* are excluded from the envelope.

ate, kaneka bedju abes, taaaa, xeiu di mikóbrius ma kriansa$_i$ *ka ta liga-, liga kela.*
Ø$_i$ *sta na brinka,*
Ø$_i$ *panha lata,*
Ø$_i$ *fazi lama,*
Ø$_i$ *brinka,*
Ø$_i$ *brinka ku areia, si na mar,*
Ø$_i$ *brinka ku areia,*
Ø$_i$ *fazi kaxtelu, kes kuza la.*

3. The above tokens were not admitted for analysis because they resemble a nonfinite clefting structure; the lack of TMA marker *ta* and the nonpast reading for the CLASS 4 verbs makes these appear to be nonfinite.

kel eeee.. ku animais també, ku otras kolegas.
Ø e normalmente asi.

4. The participant diverged on a tangent at the beginning of the Frog Story. In the next line he returns to describing the narrative briefly, but then diverges again on a similar tangent (all eligible tokens are nonetheless taken to be part of the PICTURE DESCRIPTION NARRATIVE TASK).

TIMESTAMP: 33:15
…

Kel$_i$ e unnn,

--- COUGH ---

Ø$_i$ *e un kriansa també,*

5. Subject arguments of copular *e* are excluded from the envelope.

un rapazinhu lisin, joven..
eeemm mutu filis, el ku si kãu, ku si… kasaaa…
normalmenti kriansa di Kauberdi ta kasa, ganfa- ganfanhotu,
bu konxi ganfanhoto$_i$?

--- INTERVIEWER RESPONSE ---

aaa eee [uuum insetu voador]$_i$.
tipu es e.. e ta sta mas eee, ***es***$_i$*=ta kumi palha di midju*

6. Token is excluded from the envelope because it follows a false start

--- INTERVIEWER INTERJECTION ---

Mhmm ia.
i kriansa$_j$ ta brinka ku-el$_i$,
Ø$_j$ *ta fla m-****Ø***$_i$*-eee-,*

7. SPE/anaphoric expression: Ø
   Surface form of antecedent: LEXICAL DP

   Linking: PROSODIC
   Person-number: 3PL
   Antecedent accessibility pattern: PATTERN B
   Anaphoric distance: 4
   Animacy: ANIMATE
   Specificity/definiteness: NONSPECIFIC/DEFINITE
   Turn: TURN INTERNAL
   Clause-chain position: CHAIN MEDIAL
   Clause type: MAIN

TMA: *TA*
TMA frame: NONPAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

8. Subject arguments of copular *e* are excluded from the envelope

Ø$_j$ *ta fazi ma* ***Ø***$_i$ *e di vaka,*

9. SPE/anaphoric expression: Ø
   Surface form of antecedent: ZERO

   Linking: NO LINK
   Person-number: 3PL
   Antecedent accessibility pattern: PATTERN D
   Anaphoric distance: 4
   Animacy: ANIMATE
   Specificity/definiteness: NONSPECIFIC/DEFINITE
   Turn: TURN INTERNAL
   Clause-chain position: CHAIN MEDIAL
   Clause type: MAIN
   TMA: *TA*
   TMA frame: NONPAST
   TMA frame switch: NO SWITCH
   Verb class: CLASS 4
10. Subject arguments of copular *e* are excluded from the envelope.

Ø$_j$ *ta fla ma* ***Ø***$_i$ *e vaka,*

11. Repeated forms are excluded from the envelope.
12. Subject arguments of copular *e* are excluded from the envelope.

*ma=****es***$_j$ *ta faziii vaka,*

13. Despite 11 being excluded from the envelope on the basis of repetition, it is still eligible to serve as an antecedent.
    SPE/anaphoric expression: SC
    Surface form of antecedent: ZERO

    Linking: NO LINK
    Person-number: 3PL
    Antecedent accessibility pattern: PATTERN D
    Anaphoric distance: 6
    Animacy: ANIMATE
    Specificity/definiteness: NONSPECIFIC/DEFINITE
    Turn: TURN INTERNAL
    Clause-chain position: CHAIN MEDIAL
    Clause type: MAIN
    TMA: *TA*
    TMA frame: NONPAST
    TMA frame switch: NO SWITCH
    Verb class: CLASS 4

Ø$_j$ *ta brinka ku-el*$_i$*,*

14. SPE/anaphoric expression: Ø

Surface form of antecedent: CLITIC
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 3
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

Ø$_j$ *ta, toka dianti,*

15. SPE/anaphoric expression: Ø

Surface form of antecedent: ZERO
Linking: PROSODIC
Person-number: 3PL
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 4
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

Ø$_j$ *ta pega munti* ***Ø***$_i$*,*

16. SPE/anaphoric expression: Ø

Surface form of antecedent: ZERO
Linking: PROSODIC
Person-number: 3PL
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 3
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN

TMA: *TA*
TMA frame: NONPAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

17. The latter of the two zero anaphora in this sequence is an object anaphor.

***Ø**$_j$ ta nxi na garrafa*

18. SPE/anaphoric expression: Ø

Surface form of antecedent: ZERO
Linking: PROSODIC
Person-number: 3PL
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 3
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

***Ø**$_j$ ta bira **Ø**$_j$ ta brinka ku-el.*

19. SPE/anaphoric expression: Ø

Surface form of antecedent: ZERO
Linking: PROSODIC
Person-number: 3PL
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 4
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN FINAL
Clause type: MAIN
TMA: INCORPORATED *TA* (it is the initial non-putative subject 'slot' at the head of the serial-verb-like sequence)
TMA frame: NONPAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

20. The second anaphoric zero in this sequence is an example of an 'incorporated *ta*' construction (or at least it resembles one), to the extent that it is excluded from the envelope.
21. This is the end of a clause chain ('discourse chunk').

Kes gafanhotu es=sta mas e na tenpo di azagua,

22. This is another example of a grounding clause (CLAUSE CHAIN), but does not contain an admissible token.

na tenpu di txuba,
Ki-o ki ta nasi kes palha di txon,
kes kuza ta kria mutu ti ki ***Ø*** *ta-.*

23. This anaphoric zero is exclude as a false start.

kriansa$_i$ *ta panha,*
es$_i$*=ta brinka ku-el,*

24. SPE/anaphoric expression: SC

    Surface form of antecedent: LEXICAL DP
    Linking: PROSODIC
    Person-number: 3PL
    Antecedent accessibility pattern: PATTERN B
    Anaphoric distance: 2
    Animacy: ANIMATE
    Specificity/definiteness: NONSPECIFIC/DEFINITE
    Turn: TURN INTERNAL
    Clause-chain position: CHAIN MEDIAL
    Clause type: MAIN
    TMA: *TA*
    TMA frame: NONPAST
    TMA frame switch: NO SWITCH
    Verb class: CLASS 4

Ø$_i$ *sta na rua* ***Ø***$_i$ *ta korri di sedu a tardi!*

25. SPE/anaphoric expression: Ø

    Surface form of antecedent: CLITIC
    Linking: PROSODIC
    Person-number: 3PL
    Antecedent accessibility pattern: PATTERN B
    Anaphoric distance: 4
    Animacy: ANIMATE
    Specificity/definiteness: NONSPECIFIC/DEFINITE
    Turn: TURN INTERNAL
    Clause-chain position: CHAIN MEDIAL
    Clause type: MAIN
    TMA: *BARE*
    TMA frame: NONPAST
    TMA frame switch: NO SWITCH
    Verb class: CLASS 1

26. The second anaphoric zero in this sequence is an example of an ‘incorporated *ta*’ construction (or at least it resembles one), to the extent that it is excluded from the envelope.

djuntu kesss kãu,

abes ten [kes kãu mansinhu]$_i$,
bu sabi?
***Ø**$_i$ ta sta senpri na pe di kriansa$_j$ me,*

27. This anaphoric zero is coindexed with *kes kãu mansinhu*. The interrogative or discourse marker *bu sabi?* serves as an intervening subject referent.
    SPE/anaphoric expression: Ø

    Surface form of antecedent: LEXICAL DP
    Linking: NO LINK
    Person-number: 3PL
    Antecedent accessibility pattern: PATTERN D
    Anaphoric distance: 2
    Animacy: ANIMATE
    Specificity/definiteness: NONSPECIFIC/DEFINITE
    Turn: TURN INTERNAL
    Clause-chain position: CHAIN MEDIAL
    Clause type: MAIN
    TMA: *TA*
    TMA frame: NONPAST
    TMA frame switch: NO SWITCH
    Verb class: CLASS 1

***Ø**$_{i/j/k}$ sta korri txada interu,*

28. This anaphoric zero is ambiguous; it could be coindexed with an anaphoric chain initiated by *kes kãu mansinhu,* with the lexical DP *kriansa*, or with a 3rd person plural referent involving both *kes kãu mansinhu* and *kriansa.* However, the next utterance makes clear that the referent is *kriansa.*
    SPE/anaphoric expression: Ø

    Surface form of antecedent: LEXICAL DP
    Linking: PROSODIC
    Person-number: 3PL
    Antecedent accessibility pattern: PATTERN C
    Anaphoric distance: 1
    Animacy: ANIMATE
    Specificity/definiteness: NONSPECIFIC/DEFINITE
    Turn: TURN INTERNAL
    Clause-chain position: CHAIN MEDIAL
    Clause type: MAIN
    TMA: BARE
    TMA frame: NONPAST
    TMA frame switch: X
    Verb class: CLASS 1

*abes **Ø**$_i$ ta da pankada na brasu na pe ma nem **Ø**$_i$ ka ta liga.*

29. SPE/anaphoric expression: Ø

    Surface form of antecedent: ZERO
    Linking: PROSODIC

Person-number: 3PL
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 5
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

30. SPE/anaphoric expression: Ø

Surface form of antecedent: ZERO
Linking: BOTH
Person-number: 3PL
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 9
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: COORDINATE
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: NO SWITCH
Verb class: CLASS 3

basta es e...
filisidadi ¿¿??,
e brinka, e-,
***es**$_i$=ka ten prublema na kabesa txeu.*

31. The SC *es* is excluded because of the prior false start. However, the anaphoric chain initiated by *kriansa* is continued here by *es*; it can serve as an antecedent.

*uhhh es=ka ta-, na kaza, es$_i$=ka ta liga si **Ø**$_i$ sta na kaza albes, ne?*

32. The second SC *es* is excluded on the same basis as 31, but like 31 continues the anaphoric chain.
33. SPE/anaphoric expression: Ø

Surface form of antecedent: CLITIC
Linking: BOTH
Person-number: 3PL
Antecedent accessibility pattern: PATTERN A
Anaphoric distance: 4
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE

Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: *SI*-CLAUSE
TMA: *BARE*
TMA frame: IRREALIS
TMA frame switch: SWITCH
Verb class: CLASS 1

***Ø**$_i$ sta na kaza djunto familia na brinka,*

34. SPE/anaphoric expression: *Ø*

Surface form of antecedent: ZERO
Linking: NO LINK
Person-number: 3PL
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 5
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *BARE*
TMA frame: NONPAST
TMA frame switch: SWITCH
Verb class: CLASS 1

Es e-, na rua na brinka normalmenti.

35. This ends the clause chain as well as the speakers tangential description. In the next line he returns begins to narrate the Frog Story.

*Lisin **e**$_i$=sata riprizenta klaru un rapazinhu mutu filis$_j$,*

36. This is a grounding clause that initiates a new clause chain/'discourse chunk'. The SC *e* appears to refer to the image of the boy (rather than the boy himself). Then the referent for the boy is introduced in a lexical DP *un rapazinhu mutu filis* to begin an anaphoric chain.

juntu ku si$_i$ kãu.
uuuhh pudi ser n-um jardim ou na meiu di rua.

37. The 'subject slot' for *pudi ser* is taken to be non-argumental and is excluded from the envelope.

***e**$_i$=korri korri tras di, kãu, uhhh, na brinka,*

38. SPE/anaphoric expression: SC

Surface form of antecedent: POSSESSIVE PRONOUN
Linking: NO LINK
Person-number: 3SG
Antecedent accessibility pattern: PATTERN E
Anaphoric distance: 11

Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE (although the anaphoric chain was initiated by an indefinite lexical DP, the resumption of the discourse referent on the possessive pronoun *si* in *si kau* initiates a definite referential value associated with this discourse referent's anaphoric chain).
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *BARE*
TMA frame: NONPAST (the doubling of the bare verb gives it an iterative or progressive reading, allowing it to take a NONPAST reading without necessitating a TMA marker).
TMA frame switch: X
Verb class: CLASS 4

abes ***Ø***$_i$ *ta brinka propi ku animal,*

39. SPE/anaphoric expression: Ø

Surface form of antecedent: CLITIC
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 9
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: SWITCH
Verb class: CLASS 4

abes ***Ø*** *ta atxa sapu n-um, n-um, pikenu tanki asi,*

40. This subject is generic or impersonal and does not have an antecedent lexical DP or participate in an anaphoric chain initiated by one.

albes na rua asi,
pamodi abes sapu na bem,
ali, na epoka das aguas abes,
Ø *ta atxa sapus asi alguns animais na meiu rua,*

41. Generic or impersonal (like 40).

sapu, rã$_i$*,* ***Ø*** *ta odja-s*$_i$ *albes na meiu di rua pamo na kes tenpu li,*

42. Generic or impersonal (like 40, 41).

posa di agua asi,
es$_i$*=ta kria na kes posa di agua txeu,*

43. SPE/anaphoric expression: SC

Surface form of antecedent: CLITIC
Linking: NO LINK
Person-number: 3PL
Antecedent accessibility pattern: PATTERN C
Anaphoric distance: 14
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: X
Verb class: CLASS 3

Ø *ta atxa munti animal asi,*

44. Generic or impersonal (like 40, 41, 42).

i kriansa$_i$ *ta panha Ø*$_j$*,*
abes ***Ø***$_i$ *ta brinka ku-el*$_j$*.*

45. SPE/anaphoric expression: Ø

Surface form of antecedent: LEXICAL DP
Linking: PROSODIC
Person-number: 3PL
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 3
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

abes argên grandi tem medu di Sapu ma kriansa$_i$ *eee,* ***es***$_i$ *ka mutu ten medu, aha.*

46. SCs that are coindexed with a topicalized left-dislocated lexical DP are excluded from the envelope but serves as X2SBJs when coded as antecedents to a subsequent anaphor.

o bu ta odja-l$_i$ *abes,*
Ø$_i$ *ta tropeza* ***Ø***$_i$ *ta kai,*

47. SPE/anaphoric expression: Ø

Surface form of antecedent: CLITIC
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN C

Anaphoric distance: 1
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: X
Verb class: CLASS 4

48. The second anaphoric zero in this sequence is an example of an 'incorporated *ta*' construction (or at least it resembles one), to the extent that it is excluded from the envelope.

abes ***Ø****$_i$ ta txora ma e kel un ora,*

49. SPE/anaphoric expression: Ø

Surface form of antecedent: INCORPORATED *TA* (putative 'medial' subject of an incorporated *ta* construction. These contexts were removed after constraining the envelope following Analysis 1).
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 3
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: NO SWITCH
Verb class: CLASS 3

e*$_i$=txora,*

50. Token is excluded because it is a bare CLASS 3 verb with a nonpast reading.

e*$_i$=diskansa,*

51. Token is excluded because it is a bare CLASS 4 verb with a nonpast reading.

ma dja ***Ø****$_i$ atxa brinkadera,*

52. Despite the *e* in 51 being excluded from the envelope, it can still serve as an antecedent.

SPE/anaphoric expression: Ø

Surface form of antecedent: CLITIC
Linking: BOTH
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B

Anaphoric distance: 3
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *DJA*
TMA frame: PERFECTIVE
TMA frame switch: SWITCH
Verb class: CLASS 3

kusa di torna brinka dja ***Ø****$_i$ ta skesi abes ki, tipo si,* ***Ø*** *da pankada.*

53. SPE/anaphoric expression: Ø

Surface form of antecedent: ZERO
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 7
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *DJA*
TMA frame: PERFECTIVE
TMA frame switch: NO SWITCH
Verb class: CLASS 3

54. Token is excluded because it is a bare CLASS 4 verb with a nonpast reading.

TIMESTAMP: 35:35

Ahh keli kkkk
El$_i$ e un kriansa na brinka,
i ***el****$_i$=kai na riu asi,*

1. SPE/anaphoric expression: SC

Surface form of antecedent: TONIC
Linking: BOTH
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 6
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: COORDINATE

TMA: BARE
TMA frame: PAST
TMA frame switch: SWITCH
Verb class: CLASS 4

*nem=**M**=ka parsebi si **el**$_i$=kai na riu ma=**el**$_i$ kai.*

55. This is the first mention of the 1st person subject. It can begin an anaphoric chain for subsequent 1st person anaphora.

56. SPE/anaphoric expression: SC

Surface form of antecedent: CLITIC
Linking: NO LINK
Person-number: 3SG
Antecedent accessibility pattern: PATTERN D
Anaphoric distance: 9
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: *SI*-CLAUSE
TMA: BARE
TMA frame: PAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

57. SPE/anaphoric expression: SC

Surface form of antecedent: CLITIC
Linking: BOTH
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 4
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: COORDINATE
TMA: BARE
TMA frame: PAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

*Dipôs **e**$_i$=labanta,*

58. SPE/anaphoric expression: SC

Surface form of antecedent: CLITIC
Linking: NO LINK

Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 2
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INITIAL (this in a speaker-initiated discourse chunk that was not prompted by a question.)
Clause-chain position: CHAIN INITIAL
Clause type: MAIN
TMA: BARE
TMA frame: PAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

ku rostu tristi pamodi ***Ø****$_i$ kai na agua.*

59. SPE/anaphoric expression: *Ø*

Surface form of antecedent: CLITIC
Linking: BOTH
Person-number: 3SG
Antecedent accessibility pattern: PATTERN A
Anaphoric distance: 5
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: SUBORDINATE
TMA: BARE
TMA frame: PAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

Maa ***Eeeel****$_i$ nem=****e****$_i$=ka ta liga.*

60. SPE/anaphoric expression: X2SBJ

Surface form of antecedent: ZERO
Linking: SYNTACTIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 4
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN FINAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: SWITCH
Verb class: CLASS 3

*Basta dj=**e**$_i$ kai na agua,*

61. SPE/anaphoric expression: CL

Surface form of antecedent: X2SBJ
Linking: NO LINK
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 5
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INITIAL
Clause-chain position: CHAIN INITIAL
Clause type: MAIN
TMA: *DJA*
TMA frame: PERFECTIVE
TMA frame switch: SWITCH
Verb class: CLASS 4

Ø$_i$ *ka foga nada di mal ka kontisi,*

62. SPE/anaphoric expression: Ø

Surface form of antecedent: CLITIC
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 3
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: BARE
TMA frame: PAST
TMA frame switch: SWITCH
Verb class: CLASS 3

Ø$_i$ *labanta,*

63. SPE/anaphoric expression: Ø

Surface form of antecedent: ZERO
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 7
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL

Clause type: MAIN
TMA: BARE
TMA frame: PAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

Ø$_i$ *kontinua brinka.*

64. SPE/anaphoric expression: Ø

Surface form of antecedent: ZERO
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 1
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN FINAL
Clause type: MAIN
TMA: BARE
TMA frame: PAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

Kriansa$_i$ *normalmenti e ¿¿??,*

el$_i$*=kai na agua,*

65. Token is excluded because of unidentified utterance in the second portion of the prior clause.

el$_i$*=sta na brinka,*

66. SPE/anaphoric expression: CL

Surface form of antecedent: CLITIC
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 3
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: BARE
TMA frame: PROGRESSIVE
TMA frame switch: NO SWITCH
Verb class: CLASS 1

el$_i$*=kai na agua el*$_i$ *ku si kãu,*

67. SPE/anaphoric expression: CL

Surface form of antecedent: CLITIC
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 3
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: BARE
TMA frame: PAST
TMA frame switch: SWITCH
Verb class: CLASS 4

68. Target is a singleton tonic in a coordinate lexical DP construction: DP + *ku* + DP; these forms are excluded from the envelope.

el$_i$ *ku si dogui,*

69. Target is a singleton tonic in a coordinate lexical DP construction: DP + *ku* + DP; these forms are excluded from the envelope (like 68).

ma nem Ø$_i$ *ka-, nem* ***Ø***$_i$ *ka liga,*

70. Token is excluded as following false start.

e$_i$*=kontinua ta brinka.*

71. SPE/anaphoric expression: CL

Surface form of antecedent: ZERO
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 2
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN FINAL
Clause type: MAIN
TMA: COMBINATION
TMA frame: PROGRESSIVE
TMA frame switch: SWITCH
Verb class: CLASS 4

Kuriosidadi kriansa, ali kriansa senpri e, e kuriozu.
nem ka so li,
kriansa$_i$ *senpri* ***es***$_i$ *ten kel kuriozidadi,*

72. SCs are coindexed with a topicalized left-dislocated lexical DP are excluded, but serve as antecedents.

Ø$_i$ *odja un koza novu,*

73. CLASS 3 bare verb with nonpast reading, excluded from envelope.

Ø$_i$ *prendi,*

74. CLASS 3 bare verb with nonpast reading, excluded from envelope.

*abes, iii, nem=****es***$_i$*=ka ta leba en konta ¿¿?? sima un argên mas grandi.*

75. SPE/anaphoric expression: CL

Surface form of antecedent: ZERO
Linking: PROSODIC
Person-number: 3PL
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 6
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: SWITCH
Verb class: CLASS 3

basta ***es***$_i$*=odja um kuza,*

76. SPE/anaphoric expression: CL

Surface form of antecedent: CLITIC
Linking: NO LINK
Person-number: 3PL
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 12
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: BARE
TMA frame: PAST
TMA frame switch: SWITCH
Verb class: CLASS 3

gana brinka da-s$_i$*,*

77. Post-verbal experiencer subjects of this sort were excluded from the envelope.

Ø$_i$ *bai* ***Ø***$_i$ *ta brinka dja.*

78. SPE/anaphoric expression: Ø

Surface form of antecedent: CLITIC
Linking: PROSODIC
Person-number: 3PL

Antecedent accessibility pattern: PATTERN C
Anaphoric distance: 0
Animacy: ANIMATE
Specificity/definiteness: NONSPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN FINAL
Clause type: MAIN
TMA: COMBINATION
TMA frame: PERFECTIVE
TMA frame switch: X
Verb class: CLASS 4

79. The second anaphoric element in this sequence is a putative medial subject on a incorporated *ta* sequence as is excluded from the envelope.

Lisisinli ***e***$_i$*=kai num riu di agua lisin,*

80. SPE/anaphoric expression: CL

Surface form of antecedent: CLITIC
Linking: NO LINK
Person-number: 3SG
Antecedent accessibility pattern: PATTERN G (antecedent is in a prior discourse chunk, #71)
Anaphoric distance: 58
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INITIAL
Clause-chain position: CHAIN INITIAL
Clause type: MAIN
TMA: BARE
TMA frame: PAST
TMA frame switch: X
Verb class: CLASS 4

ki podi ser istremamenti prigru pa un kriansa di kel idadi li ma..
Ø$_i$ *ka liga,*

81. SPE/anaphoric expression: Ø

Surface form of antecedent: CLITIC
Linking: NO LINK
Person-number: 3SG
Antecedent accessibility pattern: PATTERN D
Anaphoric distance: 19
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN

TMA: BARE
TMA frame: PAST
TMA frame switch: NO SWITCH
Verb class: CLASS 3

***Ø**$_i$ kai,*

82. SPE/anaphoric expression: *Ø*

Surface form of antecedent: ZERO
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 2
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: MAIN
TMA: BARE
TMA frame: PAST
TMA frame switch: NO SWITCH
Verb class: CLASS 4

***Ø** ka kontisi.*

83. It is unclear here what the referent is, though it appears to be something other than ‘the boy’. This token is excluded.

*ehhh, di akordu ku imajen lisin ka parsi m-**e**$_i$ foooga.*

84. SPE/anaphoric expression: SC

Surface form of antecedent: ZERO
Linking: NO LINK
Person-number: 3SG
Antecedent accessibility pattern: PATTERN G
Anaphoric distance: 11
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN FINAL
Clause type: SUBORDINATE
TMA: BARE
TMA frame: PAST
TMA frame switch: X
Verb class: CLASS 3

***E**$_i$ kontinua ta brinka lisisin,*

85. SPE/anaphoric expression: SC

Surface form of antecedent: CLITIC

Linking: NO LINK
Person-number: 3SG
Antecedent accessibility pattern: PATTERN G
Anaphoric distance: 11
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INITIAL
Clause-chain position: CHAIN INITIAL
Clause type: MAIN
TMA: COMBINATION
TMA frame: PROGRESSIVE
TMA frame switch: X
Verb class: CLASS 4

ki podi ser munti grandi prigu [un kriansa sozinhu]$_j$ pamodi el$_j$ so,
el$_j$ ku si$_j$ kãu,
kãu$_k$ ka pudi ten kel maturidadi asi,
kel intelijensa,
*si **Ø**$_j$ fogaba lisin pa **Ø**$_k$ djubada el$_j$.*

86. SPE/anaphoric expression: Ø

Surface form of antecedent: POSSESSIVE PRONOUN
Linking: NO LINK
Person-number: 3SG
Antecedent accessibility pattern: PATTERN F (fronted *if*-clause)
Anaphoric distance: 11
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: *SI*-CLAUSE
TMA: -*BA*
TMA frame: IRREALIS
TMA frame switch: X
Verb class: CLASS 3

87. SPE/anaphoric expression: Ø

Surface form of antecedent: LEXICAL DP
Linking: NO LINK
Person-number: 3SG
Antecedent accessibility pattern: PATTERN A
Anaphoric distance: 13
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: *PA*-CLAUSE

TMA: -*BA*
TMA frame: IRREALIS
TMA frame switch: NO SWITCH
Verb class: CLASS 4

nem [si$_j$ amigu sapu]$_l$ lisin ***Ø***$_l$ *ka debi ta djuda-l$_j$,*

88. This in an example of a DP + INTERVENING MATERIAL + Ø construction; these were not admitted to the envelope but could serve as antecedents.

Ø$_l$ *ta odja-l$_j$ ma* ***Ø***$_l$ *ka ten kel maturidadi, kel intelijensia,*

89. SPE/anaphoric expression: Ø

Surface form of antecedent: ZERO
Linking: PROSODIC
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 5
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: COORDINATE
TMA: *TA*
TMA frame: NONPAST
TMA frame switch: SWITCH
Verb class: CLASS 3

90. SPE/anaphoric expression: Ø

Surface form of antecedent: ZERO
Linking: BOTH
Person-number: 3SG
Antecedent accessibility pattern: PATTERN B
Anaphoric distance: 4
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: COORDINATE
TMA: BARE
TMA frame: NONPAST
TMA frame switch: NO SWITCH
Verb class: CLASS 1

si ***el***$_j$ *fogaba lisin,*

91. SPE/anaphoric expression: SC

Surface form of antecedent: CLITIC
Linking: NO LINK
Person-number: 3SG
Antecedent accessibility pattern: PATTERN E

Anaphoric distance: 8
Animacy: ANIMATE
Specificity/definiteness: SPECIFIC/DEFINITE
Turn: TURN INTERNAL
Clause-chain position: CHAIN MEDIAL
Clause type: *SI*-CLAUSE
TMA: -*BA*
TMA frame: IRREALIS
TMA frame switch: X
Verb class: CLASS 3

Ø*$_{l}$ bai djudaba-el$_{j}$ sai ma,*

92. Bare CLASS 4 verb with nonpast reading, excluded.

axu ki di akordu ku imajen li ka kontisi nada i kriansa asi,

93. *Axu* is an example of superstrate inflection. The inflectional suffix for these could serve as antecedents but the subject slots for these constructions were not admitted to the envelope.
94. Post-verbal subject *nada* (not relevant for envelope).

kuriosidadi abes ***Ø*** *ta leba kriansa fazi kuza ate perigrozu-*

95. Another examples of a DP + INTERVENING MATERIAL + Ø construction; these could serve as antecedents but were not admitted to the envelope.

...

Appendix 2: Analysis 1, fixed-effects multinomial logistic regression 1 - numeric statistical output

Formula = SPE ~ SURFACE FORM OF THE ANTECEDENT + LINKING + ANIMACY + PERSON/NUMBER + TMA + SESSCORE + ANTECEDENT ACCESS + CLAUSE TYPE + TASK + SPECIFICITY/DEFINITENESS + TMA FRAME SWITCH + ANAPHORIC DISTANCE + COMBINED DIALECT REGION

AIC = 3835.01
Residual Deviance = 3627.005

DV application value = SC

SURFACE FORM OF THE ANTECEDENT – application value = CLITIC

DV/p	(intercept)	DBL	LEXICAL DP	DP + INTERVENING MATERIAL + Ø	INCORPORATED *TA*
X2SBJ	-1.374	1.38	0.752	-0.13	-13.08
p	< 0.001	< 0.001	< 0.01	> 0.05	< 0.001
Ø	-1.506	-0.154	1.90	1.80	0.80
p	< 0.001	> 0.05	< 0.001	< 0.001	> 0.05

DV/p	(intercept)	INFL	POSSESSIVE PRONOUN	TONIC PRONOUN	Ø
X2SBJ	-1.374	0.314	0.09	0.714	0.511
p	< 0.001	> 0.05	> 0.05	< 0.05	> 0.05
Ø	-1.506	0.582	0.10	-0.01	1.90
p	< 0.001	> 0.05	> 0.05	> 0.05	< 0.001

LINKING – application value = NO LINK

DV/p	(intercept)	BOTH	PROSODIC	SYNTACTIC
X2SBJ	-1.374	-0.53	-1.21	-0.533
p	< 0.001	> 0.05	< 0.001	> 0.05
Ø	-1.506	0.50	0.822	-0.202
p	< 0.001	> 0.05	< 0.001	> 0.05

ANIMACY – application value = ANIMATE

DV/p	(intercept)	COLLECTIVE	INANIMATE
X2SBJ	-1.374	0.759	-11.544
p	< 0.001	> 0.05	< 0.001
Ø	-1.506	1.727	1.627
p	< 0.001	< 0.01	< 0.001

PERSON/NUMBER – application value = 1SG; ANIMACY – application value = ANIMATE

DV/p	(intercept)	1PL	3PL	3SG
X2SBJ	-1.374	-0.531	-0.912	-0.06
p	< 0.001	> 0.05	< 0.01	> 0.05
Ø	-1.506	-0.162	0.708	0.42
p	< 0.001	> 0.05	< 0.05	> 0.05

TMA – application value = *TA*

DV/p	(intercept)	AUXILIARY / MODAL	*-BA*	BARE	COMBINATION	*DJA*
X2SBJ	-1.374	-0.63	-0.40	-0.20	-0.64	-0.080
p	< 0.001	> 0.05	> 0.05	> 0.05	< 0.05	> 0.05
Ø	-1.506	-0.212	0.713	-0.45	-0.11	-0.09
p	< 0.001	> 0.05	> 0.05	< 0.01	> 0.05	> 0.05

DV/p	(intercept)	INCORPORATED *TA*	IRREGULAR	*SATA*
X2SBJ	-1.374	0.33	0.231	-2.37
p	< 0.001	> 0.05	> 0.05	< 0.001
Ø	-1.51	0.90	-0.673	-0.23
p	< 0.001	> 0.05	> 0.05	> 0.05

SOCIOECONOMIC STATUS SCORE – continuous numeric

DV/p	(intercept)	SESSCORE
X2SBJ	-1.374	-0.234
p	< 0.001	< 0.001
Ø	-1.506	-0.26
p	< 0.001	< 0.001

ANTECEDENT ACCESSIBILITY PATTERN - application value = PATTERN B

DV/p	(intercept)	PATTERN A	PATTERN C	PATTERN D	PATTERN E	PATTERN F	PATTERN G
X2SBJ	-1.374	-0.756	-0.805	0.317	-0.018	-1.284	-0.232
p	< 0.001	> 0.05	> 0.05	> 0.05	> 0.05	> 0.05	> 0.05
Ø	-1.506	-0.116	-0.837	-0.20	0.022	-0.511	-1.536
p	< 0.001	> 0.05	> 0.05	> 0.05	> 0.05	> 0.05	< 0.05

CLAUSE TYPE - application value = MAIN

DV/p	(intercept)	ADVERBIAL SUBORDINATE	COORDINATE	COPULAR COMPLEMENT	*PA* SUBORDINATE
X2SBJ	-1.374	-0.232	0.488	-12.623	-0.86
p	< 0.001	> 0.05	> 0.05	< 0.001	> 0.05
Ø	-1.506	-1.54	-0.30	-18.60	0.80
p	< 0.001	< 0.01	> 0.05	< 0.001	< 0.01

DV/p	(intercept)	QUESTION	RELATIVE	*SI*-CLAUSE	SUBORDINATE
X2SBJ	-1.374	-0.60	-1.10	-0.024	-0.615
p	< 0.001	> 0.05	> 0.05	> 0.05	> 0.05
Ø	-1.506	-1.628	-0.417	-1.044	0.63
p	< 0.001	> 0.05	< 0.01	> 0.05	< 0.01

TASK - application value = INTERVIEW

DV/p	(intercept)	FROG STORY
X2SBJ	-1.374	-0.62
p	< 0.001	< 0.05
Ø	-1.506	0.63
p	< 0.001	< 0.01

SPECIFICITY/DEFINITENESS - application value = +DEFINITE +SPECIFIC

DV/p	(intercept)	-DEFINITE -SPECIFIC	-DEFINITE +SPECIFIC	+DEFINITE -SPECIFIC
X2SBJ	-1.374	-11.05	-0.674	1.00
p	< 0.001	< 0.001	< 0.05	< 0.01
Ø	-1.506	1.27	0.244	0.582
p	< 0.001	< 0.05	< 0.05	< 0.05

TMA FRAME SWITCH - application value = NO SWITCH

DV/p	(intercept)	X (NON-SUBJECT ANTECEDENT)	SWITCH
X2SBJ	-1.374	1.44	0.33
p	< 0.001	< 0.01	< 0.05
Ø	-1.506	0.20	-0.50
p	< 0.001	< 0.05	< 0.01

ANAPHORIC DISTANCE (numeric continuous)

DV/p	(intercept)	ANAPHORIC DISTANCE
X2SBJ	-1.374	0.004
p	< 0.001	> 0.05
Ø	-1.506	-0.011
p	< 0.001	> 0.05

COMBINED DIALECT REGION - application value = SANTIAGO SUL

DV/p	(intercept)	DJARMAI	*FORA* + SUL	NORTI	SENTRU
X2SBJ	-1.374	0.50	0.70	0.12	0.132
p	< 0.001	< 0.05	< 0.001	> 0.05	> 0.05
Ø	-1.506	0.224	0.074	0.48	-0.553
p	< 0.001	> 0.05	> 0.05	> 0.05	< 0.01

Appendix 3: Analysis 2, fixed-effects multinomial logistic regression 2 - numeric statistical output

Formula = SPE ~ SURFACE FORM OF THE ANTECEDENT, LINKING, SESSCORE, ANTECEDENT ACCESS PATTERN, PERSON/NUMBER, TASK, TMA, CLAUSE TYPE, SPECIFICITY, TMA FRAME SWITCH, ANAPHORIC DISTANCE, COMBINED DIALECT REGION, AGE

Residual Deviance: 3336.629
AIC: 3520.629

DV application value = SC

SURFACE FORM OF THE ANTECEDENT – application value = CLITIC

DV/p	(intercept)	DBL	LEXICAL DP	DP + INTERVENING MATERIAL + Ø	INFL
X2SBJ	-1.60	1.41	0.80	-0.14	0.30
p	< 0.001	< 0.001	< 0.01	> 0.05	> 0.05
Ø	-1.70	-0.21	1.90	1.80	0.70
p	< 0.001	> 0.05	< 0.001	< 0.001	> 0.05

DV/p	(intercept)	POSSESSIVE PRONOUN	TONIC PRONOUN	Ø
X2SBJ	-1.60	0.03	0.70	0.53
p	< 0.001	> 0.05	< 0.05	> 0.05
Ø	-1.70	0.20	-0.053	1.90
p	< 0.001	> 0.05	> 0.05	< 0.001

LINKING – application value = NO LINK

DV/p	(intercept)	BOTH	PROSODIC	SYNTACTIC
X2SBJ	-1.60	-0.54	-1.11	-0.50
p	< 0.001	> 0.05	< 0.001	> 0.05
Ø	-1.70	0.43	0.80	-0.18
p	< 0.001	> 0.05	< 0.001	> 0.05

SOCIOECONOMIC STATUS SCORE – continuous numeric

DV/p	(intercept)	SESSCORE
X2SBJ	-1.60	-0.3
p	< 0.001	< 0.001
Ø	-1.70	-0.3
p	< 0.001	< 0.001

ANTECEDENT ACCESSIBILITY PATTERN - application value = PATTERN B

DV/p	(intercept)	PATTERN A	PATTERN C	PATTERN D	PATTERN E	PATTERN F	PATTERN G
X2SBJ	-1.60	-0.70	-0.724	0.37	0.01	-1.24	-0.26
p	< 0.001	> 0.05	> 0.05	> 0.05	> 0.05	> 0.05	> 0.05
Ø	-1.70	0.162	-0.81	-0.40	-0.40	-0.59	-1.70
p	< 0.001	> 0.05	> 0.05	> 0.05	> 0.05	> 0.05	< 0.05

PERSON/NUMBER – application value = 1SG

DV/p	(intercept)	1PL	3PL	3SG
X2SBJ	-1.60	-0.50	-1.0	-0.08
p	< 0.001	> 0.05	< 0.01	> 0.05
Ø	-1.70	-0.101	0.72	0.51
p	< 0.001	> 0.05	< 0.01	< 0.05

TASK - application value = INTERVIEW

DV/p	(intercept)	FROG STORY
X2SBJ	-1.60	-0.60
p	< 0.001	< 0.05
Ø	-1.70	0.60
p	< 0.001	< 0.01

TMA – application value = *TA*

DV/p	(intercept)	AUXILIARY / MODAL	*-BA*	BARE	COMBINATION	*DJA*
X2SBJ	-1.60	-0.641	-0.40	-0.14	-0.65	-0.06
p	< 0.001	< 0.05	> 0.05	> 0.05	< 0.05	> 0.05
Ø	-1.70	-0.321	0.70	-0.30	-0.20	-0.07
p	< 0.001	> 0.05	> 0.05	< 0.01	> 0.05	> 0.05

DV/p	(intercept)	INCORPORATED *TA*	IRREGULAR	*SATA*
X2SBJ	-1.60	0.40	0.224	-2.34
p	< 0.001	> 0.05	> 0.05	< 0.01
Ø	-1.70	1.15	-0.30	-0.03
p	< 0.001	< 0.05	> 0.05	> 0.05

CLAUSE TYPE - application value = MAIN

DV/p	(intercept)	ADVERBIAL SUBORDINATE	COORDINATE	*PA* SUBORDINATE
X2SBJ	-1.60	0.37	0.502	-0.90
p	< 0.001	> 0.05	> 0.05	< 0.01
Ø	-1.70	-1.34	-0.30	0.534
p	< 0.001	< 0.001	> 0.05	> 0.05

DV/p	(intercept)	QUESTION	RELATIVE	*SI*-CLAUSE	SUBORDINATE
X2SBJ	-1.60	-0.70	-0.53	-1.142	0.02
p	< 0.001	> 0.05	> 0.05	< 0.05	> 0.05
Ø	-1.70	-1.90	-1.42	-0.524	-1.42
p	< 0.001	< 0.01	< 0.001	> 0.05	< 0.001

SPECIFICITY - application value = +SPECIFIC

DV/p	(intercept)	-SPECIFIC
X2SBJ	-1.60	1.00
p	< 0.001	< 0.01
Ø	-1.70	0.60
p	< 0.001	< 0.05

TMA FRAME SWITCH - application value = NO SWITCH

DV/p	(intercept)	X (NON-SUBJECT ANTECEDENT)	SWITCH
X2SBJ	-1.60	1.443	0.33
p	< 0.001	< 0.05	< 0.05
Ø	-1.70	0.20	-0.55
p	< 0.001	> 0.05	< 0.001

ANAPHORIC DISTANCE – continuous numeric

DV/p	(intercept)	ANAPHORIC DISTANCE
X2SBJ	-1.60	0.004
p	< 0.001	> 0.05
Ø	-1.70	-0.01
p	< 0.001	> 0.05

AGE – continuous numeric

DV/p	(intercept)	AGE
X2SBJ	-1.60	0.01
p	< 0.001	< 0.05
Ø	-1.70	0.02
p	< 0.001	> 0.05

COMBINED DIALECT REGION - application value = SANTIAGO SUL

DV/p	(intercept)	DJARMAI	*FORA* + SUL	NORTI	SENTRU
X2SBJ	-1.60	0.473	0.70	-0.10	-0.07
p	< 0.001	< 0.05	< 0.001	> 0.05	> 0.05
Ø	-1.70	0.211	0.102	0.30	-0.80
p	< 0.001	< 0.05	> 0.05	> 0.05	< 0.001

Appendix 4: Analysis 3, mixed-effects binomial logistic regression (Model 3) - numeric statistical output

Formula = SPE ~ ANTECEDENT SURFACE + LINKING + PERSON/NUMBER + CLAUSE TYPE + SESSCORE + TMA FRAME SWITCH + TASK + (1 | PARTICIPANT)

Model AIC = 1896.3
Deviance = 1844.3
Residual Degrees of freedom = 3625

Random effects:
Variance = 0.1434
Standard Deviation = 0.3787

DV application value = OVERT

SURFACE FORM OF THE ANTECEDENT – application value = CLITIC

DV/p	(intercept)	DBL	LEXICAL DP	DP + INTERVENING MATERIAL + Ø	INFL
Ø	-2.124	-0.30	2.0	1.81	0.84
p	< 0.001	> 0.05	< 0.001	< 0.001	> 0.05

DV/p	(intercept)	POSSESSIVE PRONOUN	TONIC PRONOUN	Ø
Ø	-2.124	0.32	-0.033	1.83
p	< 0.001	> 0.05	> 0.05	< 0.001

LINKING – application value = NO LINK

DV/p	(intercept)	BOTH	PROSODIC	SYNTACTIC
Ø	-2.124	0.80	1.14	0.132
p	< 0.001	< 0.001	< 0.001	> 0.05

PERSON/NUMBER – application value = 1SG

DV/p	(intercept)	1PL	3PL	3SG
Ø	-2.124	0.30	1.11	0.70
p	< 0.001	> 0.05	< 0.001	< 0.01

CLAUSE TYPE - application value = MAIN

DV/p	(intercept)	ADVERBIAL SUBORDINATE	COORDINATE	*PA* SUBORDINATE
Ø	-2.124	-1.422	-0.411	0.602
p	< 0.001	< 0.01	> 0.05	< 0.01

DV/p	(intercept)	QUESTION	RELATIVE	*SI*-CLAUSE	SUBORDINATE
Ø	-2.124	-1.70	-1.40	-0.243	-1.434
p	< 0.001	> 0.05	< 0.05	> 0.05	< 0.001

SOCIOECONOMIC STATUS SCORE – continuous numeric

DV/p	(intercept)	SESSCORE
Ø	-2.124	-0.242
p	< 0.001	< 0.001

TMA FRAME SWITCH - application value = NO SWITCH

DV/p	(intercept)	X (NON-SUBJECT ANTECEDENT)	SWITCH
Ø	-2.124	-0.70	-0.604
p	< 0.001	< 0.01	< 0.001

TASK - application value = INTERVIEW

DV/p	(intercept)	FROG STORY
Ø	-1.70	0.40
p	< 0.001	< 0.05

Appendix 5: Analysis 4, mixed-effects binomial logistic regression (Model 4) - numeric statistical output

Formula = SPE ~ ANTECEDENT SURFACE + LINKING + PERSON/NUMBER + CLAUSE TYPE + SESSCORE + TMA FRAME SWITCH + TASK + (1 | PARTICIPANT)

Model AIC = 1679.9
Deviance = 1637.9
Residual Degrees of freedom = 3630

Random effects:
Variance = 0.2593
Standard Deviation = 0.5092

DV application value = ALL ELSE (SC + Ø)

SURFACE FORM OF THE ANTECEDENT – application value = CLITIC

DV/p	(intercept)	DBL	LEXICAL DP	DP + INTERVENING MATERIAL + Ø	INFL
X2SBJ	-2.50	1.50	0.372	-0.80	0.85
p	< 0.001	< 0.001	> 0.05	> 0.05	> 0.05

DV/p	(intercept)	POSSESSIVE PRONOUN	TONIC PRONOUN	Ø
X2SBJ	-2.50	0.30	-0.033	0.22
p	< 0.001	> 0.05	< 0.01	> 0.05

ANTECEDENT ACCESSIBILITY PATTERN - application value = PATTERN B

DV/p	(intercept)	PATTERN A	PATTERN C	PATTERN D	PATTERN E	PATTERN F	PATTERN G
X2SBJ	-1.60	-0.75	0.60	1.16	0.01	1.75	1.54
p	< 0.001	< 0.05	> 0.05	< 0.001	< 0.001	> 0.05	< 0.001

SOCIOECONOMIC STATUS SCORE – continuous numeric

DV/p	(intercept)	SESSCORE
X2SBJ	-2.50	-0.20
p	< 0.001	< 0.01

TURN - application value = MEDIAL

DV/p	(intercept)	INITIAL	RESPONSE
X2SBJ	-2.50	-0.634	-0.70
p	< 0.001	> 0.05	< 0.05

CLAUSE CHAIN POSITION - application value = CHAIN MEDIAL

DV/p	(intercept)	CHAIN FINAL	GROUNDING CLAUSE	CHAIN INITIAL
X2SBJ	-2.50	-0.50	-0.65	0.70
p	< 0.001	> 0.05	> 0.05	< 0.01

Bibliography

Abreu, Laurel. 2012. Subject Pronoun Expression and Priming Effects among Bilingual Speakers of Puerto Rican Spanish. *Selected Proceedings of the 14th Hispanic Linguistics Symposium* (Cascadilla Proceedings Project), 1–8. Cascadilla.

Adams, Marianne Patalino. 1987. *Old French, Null Subjects, and Verb Second Phenomena: A Dissertation.* University of California, Los Angeles. (Doctoral dissertation.)

Adger, David & Jennifer Smith. 2005. Variation and the minimalist program. In Leonie M.E.A. Cornips & Karen P. Corrigan (eds.), *Current Issues in Linguistic Theory*, vol. 265, 149–178. Amsterdam: John Benjamins Publishing Company.

Alonso, Vicenta Cortés. 1964. *La esclavitud en Valencia durante el reinado de los Reyes Católicos (1479-1516)*. Valencia: Excmo. Ayuntamiento.

Alonso-Ovalle, Luis, Susana Fernández-solera & Lyn Frazier. 2002. Null vs. overt pronouns and the topic-focus articulation in Spanish. *Journal of Italian Linguistics* 14(2). 151–170.

Amado, Abel Djassi. 2015. The illegible state in Cape Verde: language policy and the quality of democracy.

Andrade, Aroldo. 2015. On the Emergence of Topicalisation in European Portuguese: a study at the Syntax-Information Structure Interface. *Estudos Linguísticos* 11.

Andrade, Elisa Silva. 1996. *Les îles du Cap-Vert de la "Découverte" à l'indépendance nationale: 1460-1975*. Paris: L'Harmattan.

Antonelli, André Luís. 2011. *Sintaxe da posição do verbo e mudança gramatical na história do português Europeu*. Universidade Estadual de Campinas Text. (Doctoral dissertation.)

Arends, Jacques. 1993. Towards a Gradualist Model of Creolization. In Francis Byrne & John Holm (eds.), *Atlantic Meets Pacific: A global view of pidginization and creolization* (Creole Language Library 11). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Arends, Jacques. 2000. A Demographic Perspective on Creole Formation. In Silvia Kouwenberg & John Victor Singler (eds.), *The Handbook of Pidgin and Creole Studies*. Hoboken, UK: John Wiley & Sons.

Ariel, Mira. 1993. *Accessing Noun-Phrase Antecedents* (RLE Linguistics B: Grammar). Routledge.

Ariel, Mira. 1999. The development of person agreement markers: From pronouns to higher Accessibility markers. In Michael Barlow & Suzanne Kemmer (eds.), *Usage based Models of Language*, 197–260. CSLI.

Ariel, Mira. 2001. Accessibility theory: An overview. In Ted J.M. Sanders, Joost Schilperoord & Wilbert Spooren (eds.), *Text representation: Linguistic and psycholinguistic aspects* (Human Cognitive Processing 8), 29–87. John Benjamins Publishing Company.

Armstrong, Nigel. 2002. Variable deletion of French ne: a cross-stylistic perspective. *Language Sciences* 2(24). 153–173.

Ashby, William J. 1980. Prefixed conjugation in Parisian French. In Herbert J. Izzo (ed.), *Italic and Romance: Linguistic studies in honor of Ernst Pulgram* (Current Issues in Linguistic Theory 18). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Ashby, William J. 1982. The drift of French syntax. *Lingua* 57(1). 29–46.

Auger, Julie. 1994. Pronominal Clitics in Québec Colloquial French: A Morphological Analysis. *IRCS Technical Reports Series*. https://repository.upenn.edu/ircs_reports/174.

Baayen, R. Harald. 2008. *Analyzing linguistic data: a practical introduction to statistics using R*.

Cambridge, UK ; New York: Cambridge University Press.
Bailey, Guy & Jan Tillery. 2004. Some sources of divergent data. In Ronald K. S. Macaulay & Carmen Fought (eds.), *Sociolinguistic Variation: Critical Reflections*. Oxford University Press.
Bailey, Laura. 2011. Null subjects in Northeast English. *Newcastle working papers in linguistics*, vol. 17, 23–45.
Baker, B. 2002. How referential is agreement?: The interpretation of polysynthetic dis-agreement morphology in Ngalakgan. *Problems of Polysynthesis*, 51–86. Berlin: Akademie Verlag.
Baker, Philip. 1972. *Kreol: a description of Mauritian Creole*. London: C. Hurst.
Bakker, Peter. 2009. Pidgins versus Creoles and Pidgincreoles. *The Handbook of Pidgin and Creole Studies*, 130–157. John Wiley & Sons, Ltd.
Baleno, Ilídio. 2007a. Povoamento e formação da sociedade. In Maria Emília Madeira Santos, Maria Manuel Torrão & Maria João Soares (eds.), *História concisa de Cabo Verde*, vol. 1. 2nd edn. Instituto de Investigação Científica Tropical.
Baleno, Ilídio. 2007b. Pressões externas: reacções ao corso e à pirataria. In Maria Emília Madeira Santos, Maria Manuel Torrão & Maria João Soares (eds.), *História concisa de Cabo Verde*, vol. 2. 2nd edn. Instituto de Investigação Científica Tropical.
Baleno, Ilídio. 2007c. Reconversão do comércio externo em tempo de crise e o impacto da companhia do Grão-Pará e Maranhão. In Maria Emília Madeira Santos, Maria Manuel Torrão & Maria João Soares (eds.), *História concisa de Cabo Verde*, vol. 3. 2nd edn. Instituto de Investigação Científica Tropical.
Baptista, Marlyse. 1995. On the pro-drop parameter in Cape Verdean Creole. *Harvard Working Papers in Linguistics* 5. 3–17.
Baptista, Marlyse. 1997. From Orality to the Written Word: A Linguistic Rite of Passage. *Cimboa* 2(2). 7.
Baptista, Marlyse. 2002. *The syntax of Cape Verdean Creole: the Sotavento varieties* (Linguistik Aktuell/Linguistics Today v. 54). Amsterdam/Philadelphia: John Benjamins Publishing.
Baptista, Marlyse. 2004. A cross linguistic comparison of copula predication: some basic assumptions revisited. *Journal of Portuguese Linguistics* 3. 97-113.
Baptista, Marlyse. 2006. Selection and Competition in Creole Formation: A Case Study. In Jürgen Lang, John Holm & Jean-Louis Rougé (eds.), *Cabo Verde - origens da sua sociedade e do seu crioulo*. Tübingen: Gunter Narr Verlag.
Baptista, Marlyse. 2007. Selection and Competition in Creole Formation: A Case Study. *University of Pennsylvania Working Papers in Linguistics*, vol. 2, 14. University of Pennsylvania Libraries.
Baptista, Marlyse, Heliana Mello & Miki Suzuki. 2007. Kabuverdianu, or Cape Verdean, and Kriyol, or Guinea-Bissau (Creole Portuguese. In John A. Holm & Peter L. Patrick (eds.), *Comparative Creole syntax: parallel outlines of 18 Creole grammars* (Westminster Creolistics Series 7), 53–82. London: Battlebridge.
Barbosa, Pilar, Maria Eugênia L. Duarte & Mary Aizawa Kato. 2005. Null Subjects in European and Brazilian Portuguese. *Journal of Portuguese Linguistics* 4(2). 11.
Barry, Boubacar. 1998. *Senegambia and the Atlantic Slave Trade*. Cambridge University Press.
Bartens, Ángela. 2003. *A contrastive grammar: Islander - Caribbean Standard English - Spanish*. Finnish Academy of Science and Letters.
Bartens, Ángela. 2006. A contribuição do substrato africano para a génese dos crioulos caboverdianos: o caso dos ideofones. In Jürgen Lang, John Holm & Jean-Louis Rougé

(eds.), *Cabo Verde - origens da sua sociedade e do seu crioulo*. Tübingen: Gunter Narr Verlag.
Bartens, Ángela & Eeva Sippola. 2014. Subject null arguments in creole languages. *Time and Space in Linguistics: Interdisciplinary Computational Approaches & Cross-Creole Comparison.* Aarhus University.
Batalha, Luís. 2004. *The Cape Verdean diaspora in Portugal: colonial subjects in a postcolonial world*. Lanham, MD: Lexington Books.
Batalha, Luís & Jørgen Carling (eds.). 2008. *Transnational archipelago: perspectives on Cape Verdean migration and diaspora*. Amsterdam: Amsterdam University Press.
Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using `lme4`. *Journal of Statistical Software* 67(1).
Belsley, David A. 1991. *Collinearity Diagnostics: Collinearity and Weak Data in Regression*. New York: John Wiley.
Benincà, Paola. 1984. Un'ipotesi sulla sintassi delle lingue romanze medievali. *Quaderni Patavini di Linguistica* 4(3).
Benincà, Paola. 2014. "Subject" clitics and particles in Provençal. *Probus* 26(2). 183–215.
Bentivoglio, Paola. 1987. *Los sujetos pronominales de primera persona en el habla de Caracas*. Caracas: Universidad Central de Venezuela.
Besten, Hans den. 1983. On the interaction of root transformations and lexical deletive rules. In Werner Abraham (ed.), *On the Formal Syntax of the Westgermania*, 47–131. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Bickerton, Derek. 1975. *Dynamics of a creole system*. London ; New York: Cambridge University Press.
Bickerton, Derek. 2001. Okay for content words, but what about functional items? *The Behavioral and Brain Sciences* 24(6). 1104–1105.
Bigman, Laura. 1993. *History and Hunger in West Africa: Food Production and Entitlement in Guinea-Bissau and Cape Verde*. Greenwood Press.
Bock, J. Kathryn. 1986. Syntactic persistence in language production. *Cognitive Psychology* 18(3). 355–387.
Bouchard, Marie-Eve. 2018. Subject Pronoun Expression in Santomean Portuguese. *Journal of Portuguese Linguistics* 17. 5.
Boxer, C. R. 1991. *The Portuguese seaborne empire 1415-1825* (Aspects of Portugal). 2nd ed. Manchester: Carcanet in association with the Calouste Gulbenkian Foundation.
Brandi, Luciana & Patrizia Cordin. 1989. Two Italian Dialects and the Null Subject Parameter. In Osvaldo A. Jaeggli & Kenneth J. Safir (eds.), *The Null Subject Parameter*, vol. 15, 111–142. Dordrecht: Springer Netherlands.
Brásio, P.A.D. 1962. *Descobrimento, povoamento, evangelização do arquipélago de Cabo Verde*. Lisboa: Centro de Estudos Históricos Ultramarinos.
Bresnan, Joan & Sam A. Mchombo. 1987. Topic, Pronoun, and Agreement in Chichewa. *Language* 63(4).
Brook, Marisa. 2018. Taking it up a level: Copy-raising and cascaded tiers of morphosyntactic change. *Language Variation and Change* 30(2). 231–260.
Brooks, George E. 2003. *Eurafricans in western Africa: commerce, social status, gender, and religious observance from the sixteenth to the eighteenth century*. Athens: Ohio University Press.
Buchstaller, Isabelle. 2009. The Quantitative Analysis of Morphosyntactic Variation: Constructing

and Quantifying the Denominator. *Language and Linguistics Compass* 3(4).
Bullock, Barbara E. & Almeida Jacqueline Toribio. 2009. Reconsidering Dominican Spanish: Data from the rural Cibao. *Revista Internacional de Lingüística Iberoamericana* 7(2). 49–73.
Bybee, Joan, Revere Perkins & William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. University of Chicago Press.
Cabral, Iva Maria. 2007b. Política e sociedade: ascensão e queda de uma elite endogena. In Maria Emília Madeira Santos, Maria Manuel Torrão & Maria João Soares (eds.), *História concisa de Cabo Verde*, vol. 3. 2nd edn. Instituto de Investigação Científica Tropical.
Calabrese, A. 1986. Pronomina: Some Properties of the Italian Pronominal System. In N Fukui, T. R Rapoport & E Sagey (eds.), *Papers in Theoretical Linguistics* (MIT Working Papers in Linguistics 8).
Camacho, José. 2008. Syntactic Variation: The Case of Spanish and Portuguese Subjects. *Studies in Hispanic and Lusophone Linguistics* 1(2). 415–434.
Camacho, José. 2013. *Null subjects* (Cambridge Studies in Linguistics 137). Cambridge ; New York: Cambridge University Press.
Camacho, José. 2016. The Null Subject Parameter revisited. The evolution from null subject Spanish and Portuguese to Dominican Spanish and Brazilian Portuguese. In Maru Kato & Francisco Ordróñez (eds.), *Morphosyntax of Portuguese and Spanish in Latin America*, 27–48.
Camacho, José & José Elías-Ulloa. 2010. Null subjects in Shipibo switch-reference systems. *Information Structure in Indigenous Languages of the Americas. Syntactic Approaches*, 65–86. Berlin, New York: De Gruyter Mouton.
Cameron, Richard. 1993. Ambiguous agreement, functional compensation, and non-specifictúin the Spanishof San Juan, Puerto Rico, and Madrid, Spain. *Language Variation and Change* 5. 305–334.
Cameron, Richard & Nydia Flores-Ferrán. 2004. Perseveration of subject expression across regional dialects of Spanish. *Spanish in Context* 1(1). 41–65. doi:10.1075/sic.1.1.05cam.
Cardinaletti, Anna & Michal Starke. 1994. The Typology of Structural Deficiency : on the three grammatical classes. *University of Venice Working Papers in Linguistics* 4(2). 41–109.
Cardinaletti, Anna & Michal Starke. 1996. Deficient pronouns: A view from Germanic. In Epstein Thráinsson & S Peter (eds.), *Studies in Comparative Germanic Syntax*, vol. 2. Kluwer.
Cardinaletti, Anna & Michal Starke. 1999. The typology of structural deficiency: A case study of the three classes of pronouns. In Henk van Riemsdijk (ed.), *Eurotyp*. Berlin/New York: De Gruyter Mouton
Carminati, Maria Nella. 2002. *The processing of Italian subject pronouns*. University of Massachusetts at Amherst. (Doctoral dissertation.)
Carreira, António. 1972. *Cabo Verde: formação e extinção de uma sociedade escravocrata (1460-1878).* Centro de Estudos da Guiné Portuguesa.
Carreira, António. 1982. *The People of the Cape Verde Islands: Exploitation and Emigration*. C. Hurst.
Carreira, António. 1983. *Migrações nas Ilhas de Cabo-Verde*. Instituto Caboverdeano do Livro.
Carter, Katherine & Judy Aulette. 2009. Creole in Cape Verde: Language, identity and power. *Ethnography* 10(2). 213–236.
Carvalho, Ana M. & Ryan M. Bessett. 2015. Subject pronoun expression in Spanish in contact with Portuguese. In Ana Maria Carvalho, Rafael Orozco & Naomi Lapidus Shin (eds.), *Subject Pronoun Expression in Spanish. A Cross-dialectal Perspective*, 143–168.

Carvalho, Ana M & Michael Child. 2011. Subject Pronoun Expression in a Variety of Spanish in Contact with Portuguese. In Jim Michnowicz & Robin Dodsworth (eds.), *Selected Proceedings of the 5th Workshop on Spanish Sociolinguistics*, 14–25. Cascadilla Proceedings Project.

Carvalho, Ana Maria, Rafael Orozco & Naomi Lapidus Shin. 2015. Introduction. *Subject pronoun expression in Spanish: a cross-dialectal perspective* (Georgetown Studies in Spanish Linguistics). Georgetown University Press.

Carvalho, Ana Maria, Rafael Orozco & Naomi Lapidus Shin (eds.). 2015. *Subject pronoun expression in Spanish: a cross-dialectal perspective* (Georgetown Studies in Spanish Linguistics). Washington, D.C: Georgetown University Press.

Castro, Ivo. 2006. *Introdução à História do Português*. 2nd edn. Edições Colibri.

Cedergren, Henrietta J. & David Sankoff. 1974. Variable Rules: Performance as a Statistical Reflection of Competence. *Language* 50(2). 333.

Cerrón-Palomino, Álvaro. 2018. Null-subject encounter: Variable subject pronoun expression in the Spanish of Quechua-Spanish bilinguals in the Central Peruvian Andes. *International Journal of Bilingualism* 23(5). 1005–1023.

Chafe, Wallace. 1987. Cognitive constraints on information flow. In Russel S. Tomlin (ed.), *Coherence and Grounding in Discourse: Outcome of a Symposium, Eugene, Oregon, June 1984* (Typological Studies in Langauge 11), 21–51.

Chafe, Wallace L. 1993. Prosodic and functional units of language. In Jane A. Edwards & Martin D. Lampert (eds.), *Talking data. Transcription and coding in discourse research.* Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Chafe, Wallace L. 1994. *Discourse, consciousness, and time: the flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

Chang, Franklin, Gary S. Dell, Kathryn Bock & Zenzi M. Griffin. 2000. Structural Priming as Implicit Learning: A Comparison of Models of Sentence Production. *Journal of Psycholinguistic Research* 29(2). 217–230.

Chaudenson, Robert. 1992. Des îles, des hommes, des langues: Langues créoles, cultures créoles. Paris: L'Harmattan

Chaudenson, Robert. 2001. *Creolization of Language and Culture*. Routledge.

Cheshire, Jennifer. 1987. Syntactic variation, the linguistic variable and sociolinguistic theory. *Linguistics* 25. 257–282.

Childs, Claire. 2017. Variation and Change in English Negation: A Cross-Dialectal Perspective. 328. University of New Castle upon Tyne. (Doctoral dissertation.)

Chomsky, Noam. 1981. *Lectures on government and binding: The Pisa lectures* (Studies in Generative Grammar). 7th ed. Berlin ; New York: Mouton de Gruyter.

Chomsky, Noam & Howard Lasnik. 1990. Filters and Control. *Essays on Restrictiveness and Learnability*, vol. 20, 42–124. Dordrecht: Springer Netherlands.

Cintra, Luis F. Lindley. 1986. *Sobre Formas de Tratamento na Língua Portuguesa*. 2nd edn. Lisboa: Livros Horizonte.

Claes, Jeroen. 2017. Probabilistic Grammar: The view from Cognitive Sociolinguistics. *Glossa: a journal of general linguistics* 2(1). 62.

Cohen, Zelinda. 2007b. A Administração das Ilhas de Cabo Verde Pós Uniao IMrica: continuidades e rupturas. In Maria Emília Madeira Santos, Maria Manuel Torrão & Maria João Soares (eds.), *História concisa de Cabo Verde*, vol. 3. 2nd edn. Instituto de Investigação Científica Tropical.

Cole, Melvyn Douglas. 2009. Null subjects: a reanalysis of the data. *Linguistics* 47(3).
Coonan, Patrick J. 2007. *The Language Debate in Cape Verde*. Ohio University.
Cornips, Leonie M.E.A. & Karen P. Corrigan. 2005. Toward an integrated approach to syntactic variation: A retrospective and prospective synopsis. In Leonie M.E.A. Cornips & Karen P. Corrigan (eds.), *Current Issues in Linguistic Theory*, vol. 265, 149–178. Amsterdam: John Benjamins Publishing Company.
Correia e Silva, António. 2007a. Espaço, ecologia e economia interna. In Maria Emília Madeira Santos, Maria Manuel Torrão & Maria João Soares (eds.), *História concisa de Cabo Verde*, vol. 1. 2nd edn. Instituto de Investigação Científica Tropical.
Correia e Silva, António. 2007b. Dinâmicas de decomposição recomposição dos espaços e sociedades. In Maria Emília Madeira Santos, Maria Manuel Torrão & Maria João Soares (eds.), *História concisa de Cabo Verde*, vol. 2. 2nd edn. Instituto de Investigação Científica Tropical.
Correia e Silva, António. 2007c. Reconversão do comércio externo em tempo de crise e o impacto da companhia do Grão-Pará e Maranhão. In Maria Emília Madeira Santos, Maria Manuel Torrão & Maria João Soares (eds.), *História concisa de Cabo Verde*, vol. 3. 2nd edn. Instituto de Investigação Científica Tropical.
Costa, João & Fernanda Pratas. 2008. Licenciar pro não significa ser uma língua pro-drop: evidência do caboverdiano. In Sónia Frota & Ana Lúcia Santos (eds.), *XXIII Encontro Nacional da Associação Portuguesa de Linguística*, 157–166.
Costa, João & Fernanda Pratas. 2013. Embedded null subjects in Capeverdean. *Journal of Linguistics* 49(1). 33–53.
Coupland, Nikolas. 1980. Style-shifting in a Cardiff work-setting. *Language in Society* 9(1). 1–12.
Coveney, Aidan. 2002. *Variability in spoken French: a sociolinguistic study of interrogation and negation*. Exeter: Elm Bank.
Cowper, Elizabeth. 2002. *Finiteness*. University of Toronto.
Creissels, Denis. 2010. The flexibility of the noun vs. verb distinction in the lexicon of Mandinka. *International Conference on Polycategoriality.* Paris, October 4-6, 2010.
Creissels, Denis. 2005. A typology of subject marker and object marker systems in African languages. In Voeltz F.K.E. (ed.), *Studies in African Linguistic Typology*. Amsterdam/Philadelphia: John Benjamins Publishing Company**.**
Creissels, Denis. 2015. R-impersonals in Atlantic and Mande languages. *Colloque Pronoms R-impersonnels.* Paris, February 4-6, 2015.
Creissels, Denis. 2017. The flexibility of the noun/verb distinction in the lexicon of Mandinka. In Valentina Vapnarsky & Edy Veneziano (eds.), *Lexical Polycategoriality: Cross-linguistic, cross-theoretical and language acquisition approaches* (Studies in Language Companion Series 182). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Croissant, Yves. 2010. *Estimation of multinomial logit models in R : The mlogit Packages*. https://pdfs.semanticscholar.org/8d40/143d338c298a4b5e6d421a730d54908c9eba.pdf.
Cukor-Avila, Patricia & Guy Bailey. 2001. The effects of the race of the interviewer on sociolinguistic fieldwork. *Journal of Sociolinguistics* 5(2). 252–270.
Culbertson, Jennifer. 2010. Convergent evidence for categorial change in French: From subject clitic to agreement marker. *Language* 86(1). 85–132.
Culbertson, Jennifer & Simon Kirby. 2016. Simplicity and Specificity in Language: Domain-General Biases Have Domain-Specific Effects. *Frontiers in Psychology* 6.

Curtin, Philip D. 1975. *Economic change in precolonial Africa: Senegambia in the era of the slave trade*. University of Wisconsin Press.
Dalton, Heather. 2012. 'Into speyne to selle for slavys': English, Spanish, and Genoese Merchant Networks and their Involvement with the 'Cost of Gwynea' Trade before 1550. In Tobias Green (ed.), *Brokers of Change: Atlantic Commerce and Cultures in Pre-Colonial Western Africa*. British Academy.
Davidson, Brad. 1996. 'Pragmatic weight' and Spanish subject pronouns: The pragmatic and discourse uses of 'tú' and 'yo' in spoken Madrid Spanish. *Journal of Pragmatics* 26(4). 543–565.
Davydov, Artem. 2010. Historical Morphology of Personal Pronouns in Manding. In Konstantin Pozdniakov, Valentin Vydrin & Alexander Zheltov (eds.), *Personal pronouns in Niger-Congo languages: International workshop. St. Petersburg, September 13-15, 2010. Abstracts and papers: Working materials.*, 23–37. St. Petersburg University Press.
De Cat, Cécile. 2005. French subject clitics are not agreement markers. *Lingua* 115(9). 1195–1219.
De Cat, Cécile. 2007. *French Dislocation: Interpretation, Syntax, Acquisition*. Oxford University Press.
DeGraff, Michel. 1993. Is Haitian Creole a Pro-Drop Language? In Francis Byrne & John Holm (eds.), *Creole Language Library*, vol. 11, 71. Amsterdam: John Benjamins Publishing Company.
Déprez, Viviane. 1994. Haitian Creole: A Pro-Drop Language? *Journal of Pidgin and Creole Languages* 9(1). 1–24.
Dias, Hildizina. 2009. O parâmetro do sujeito nulo em aprendentes do português em Moçambique. In Hildizina Dias (ed.), *Português moçambicano: Estudos e reflexões*. Maputo: Imprensa Universitária.
Dione, Cheikh Bamba. 2013. Handling Wolof clitics in LFG. In Christine Meklenborg Salvesen & Hans Petter Helland (eds.), *Challenging Clitics* (Linguistik Aktuell/Linguistics Today), vol. 206, 87–118. Amsterdam: John Benjamins Publishing Company.
Do Couto. 2003. Comunidade e linguagem na costa da Guiné nos séculos XV, XVI e XVII. *Polifonia* 7. 87–102.
Do Couto, Hildo Honório. 1992. Lançados, grumetes e a origem do crioulo português no noroeste africano. In Ernesto d'Andrade & Alain Kihm (eds.), *Actas do do Colóquio sobre "Crioulos de base lexical portuguesa,"* 109–122. Lisboa: Edições Colibri.
Do Couto, Hildo Honório & U Souza. 2006. As consoantes pré-nasalizadas no crioulo caboverdiano: por uma interpretação bifonemática. In Jürgen Lang, John Holm & Jean-Louis Rougé (eds.), *Cabo Verde - origens da sua sociedade e do seu crioulo*. Tübingen: Gunter Narr Verlag.
Domingues, Ângela. 2007. Administração e instituições: transplante, adaptação, funcionamento. In Maria Emília Madeira Santos, Maria Manuel Torrão & Maria João Soares (eds.), *História concisa de Cabo Verde*, vol. 1. 2nd edn. Instituto de Investigação Científica Tropical.
Donelha, André. 1977. *Descrição da Serra Leoa e dos rios de Guiné do Cabo Verde,1625 =: An account of Sierra Leone and the rivers of Guinea of Cape Verde, 1625* ([Série de memórias] - Centro de Estudos de Cartografia Antiga, Secção de Lisboa; 19). (ed.) P.E.H. Hair. Lisboa: Junta de Investigações Científicas do Ultramar.
Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming & Danae Paolino. 1993. Outline of discourse transcription. In Jane A. Edwards & Martin D. Lampert (eds.), *Talking data:*

Transcription and coding in discourse research, 45–89. Hillsdale, NJ: Lawrence Erlbaum.
Duarte, Maria Dulce de Oliveira Almada. 1998. *Bilinguismo ou diglossia?: as relaçaões de força entre o crioulo e o português na sociedade cabo-verdiana*. Spleen Ed. University of Michigan.
Duarte, Maria Eugênia L. 1993. Do pronome nulo ao pronome pleno: a trajetória do sujeito no português do Brasil. In Ian Roberts & Mary Aizawa Kato (eds.), *Português Brasileiro: uma viagem diacrônica.*, 107–128. Campinas: SP: Editorial da UNICAMP.
Duarte, Maria Eugênia L. 1995. *A Perda do Princípio "Evite pronome" no Português Brasileiro*. UNICAMP. (Doctoral dissertation.)
Duarte, Maria Eugênia L. 2000. The loss of the 'Avoid Pronoun' principle in Brazilian Portuguese. In Mary Aizawa Kato & Esmeralda Vailati Negrão (eds.), *Brazilian Portuguese and the Null Subject Parameter* (Ediciones de Iberoamericana 4). Frankfurt am Main, Madrid: Iberoamericana Vervuert.
Duarte, Maria Eugênia L. 2004. On the 'embedding' of a syntactic change. In Britt-Louise Gunnarsson & International Conference on Language Variation in Europe (eds.), *Language variation in Europe: papers from the 2nd International Conference on Language Variation in Europe, ICLaVE 2; June 12-14, 2003*, 145–155. Uppsala, Sweden: Uppsala University, Sweden.
Duarte, Maria Eugênia L. & Mary Aizawa Kato 2014. Variation in syntax: Null expletives and raised constituents in Brazilian Portuguese. *43rd New Ways of Analyzing Variation.* Chicago.
Duarte, Maria Eugênia L. & Humberto Soares da Silva. 2016. Microparametric variation in Spanish and Portuguese. In Mary A. Kato & Francisco Ordoñez (eds.), *The Morphosyntax of Portuguese and Spanish in Latin America*, 1–26. Oxford University Press.
Duffy, James. 1962. *Portugal in Africa*. Cambridge: Harvard University Press.
Dunigan, Melynda B. 1994. *On the Clausal Structure of Wolof*. University of North Carolina at Chapel Hill.
Eberhard, David M., Gary F. Simons & Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the World. *Ethnologue*. Dallas, TX: SIL International.
Estrada-Fernandez, Zarina. 2016. Exploring finiteness and non-finiteness in Pima Bajo (Uto-Aztecan). In Claudine Chamoreau & Zarina Estrada-Fernandez (eds.), *Finiteness and Nominalization* (Typological Studies in Language 113). https://benjamins.com/catalog/tsl.113.03est (19 November, 2019).
Farrington, Charles Ryan. 2019. *Language Variation and Change: Regionality and African American Language*. University of Oregon. (Doctoral dissertation.)
Flores-Ferrán, Nydia. 2002. *Subject personal pronouns in Spanish narratives of Puerto Ricans in New York City: a sociolinguistic perspective* (LINCOM Studies in Sociolinguistics 02). Muenchen: LINCOM Europa.
Foley, William A. & Robert D. Van Valin Jr. 1984. *Functional Syntax and Universal Grammar*. Cambridge University Press.
Fonseca-Greber, Bonnibeth. 2000. *The Change from pronoun to clitic to prefix and the rise of null subjects in spoken Swiss French*. University of Arizona.
Galves, Charlotte. 1996. Clitic-placement and parametric change in Portuguese. In Mario Saltarelli (ed.), *Aspects of Romance linguistics: Selected papers from the 24th LSRL*, 227–40. Washington, D.C.: Georgetown University Press.
Galves, Charlotte, Helena Britto & Maria Clara Paixão de Sousa. 2005. The Change in Clitic

Placement from Classical to Modern European Portuguese: Results from the Tycho Brahe Corpus. *Journal of Portuguese Linguistics* 4(1). 39–67. doi:10.5334/jpl.166.
Galves, Charlotte & Maria Clara Paixão de Sousa. 2017. The change in the position of the verb in the history of Portuguese: Subject realization, clitic placement, and prosody. *Language* 93(3). 152–180.
Gilligan, Gary Martin. 1987. *A cross-linguistic approach to the pro-drop parameter*. University of Southern California. (Doctoral dissertation.)
Givón, Talmy. 1971. Historical syntax and synchronic morphology: An archaeologist's field trip. *Papers from the 7th Meeting of the Chicago Linguistic Society*. Chicago Linguistics Society.
Givón, Talmy. 1976. Topic, pronoun and grammatical agreement. In Charles N. Li (ed.), *Subject and Topic*, 151–188. New York: Academic Press.
Givón, Talmy. 1983a. Topic continuity in discourse: The functional domain of switch-reference. In John Haiman & Pamela Munro (eds.), *Switch Reference and Universal Grammar: Proceedings of a symposium on switch reference and universal grammar, Winnipeg, May 1981* (Typological Studies in Langauge 2). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Givón, Talmy. 1983b. Topic continuity in spoken English. In Talmy Givón (ed.), *Topic continuity in discourse: A quantitative cross-language study* (Typological Studies in Language 3). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Givón, Talmy. 1983c. An introduction. In Talmy Givón (ed.), *Topic continuity in discourse: A quantitative cross-language study* (Typological Studies in Language 3). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Givón, Talmy. 1983d. Topic continuity and word-order pragmatics in Ute. In Talmy Givón (ed.), *Topic continuity in discourse: A quantitative cross-language study* (Typological Studies in Language 3). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Givón, Talmy. 1992. The grammar of referential coherence as mental processing instructions. In Norbert Dittmar (ed.), *Topic: special issue of Linguistics*, vol. 30, 5–56. 1st edn.
Givón, Talmy. 2001. *Syntax: An Introduction*. 2nd edn. John Benjamins Publishing Company.
Givón, Talmy. 2017. *The Story of Zero*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Givón, Talmy. 2018. *On Understanding Grammar*. 2nd edn. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Goldberg, Adele E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.
Goria, Cecilia. 2004. *Subject Clitics in the Northern Italian Dialects: a Comparative Study Based on the Minimalist Program and Optimality Theory.* Dordrecht: Springer Netherlands.
Green, Lisa J. 2002. *African American English: A Linguistic Introduction*. Cambridge University Press.
Green, Tobias. 2006. *Masters of Difference: Creolization and the Jewish Presence in Cabo Verde, 1497-1672*. University of Birmingham.
Green, Toby. 2009. Architects of Knowledge, Builders of Power: Constructing the Kaabu "Empre", 16th - 17th centuries. *Mande Studies* 11. 91–112.
Gumperz, John J. & Norine B. Berenz. 1993. Transcribing conversational exchanges. In Jane A. Edwards & Martin D. Lampert (eds.), *Talking data: Transcription and coding in discourse research*, 91–121. Hillsdale, NJ: Lawrence Erlbaum.

Gutiérrez-Rexach, Javier & Sandro Sessarego. 2011. The valuation of gender agreement in DP. In Janine Berns, Haike Jacobs & Tobais Scheer (eds.), *Romance Languages and Linguistic Theory 2009: Selected papers from "Going Romance" Nice 2009* (Romance Languages and Linguistic Theory 3), 133–148.

Gutiérrez-Rexach, Javier & Sandro Sessarego. 2014a. Morphosyntactic variation and gender agreement in three Afro-Andean dialects. *Lingua* (Morphological Variation in Spanish) 151. 142–161.

Gutiérrez-Rexach, Javier & Sandro Sessarego. 2014b. In Rafael Orozco (ed.), *New Directions in Hispanic Linguistics*, 188–216. Newcastle: Cambridge Scholars Publishing.

Hadfield, Jarrod D. 2010. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software* 33(2). 1–22.

Haegeman, Liliane. 1990. Understood subjects in English diaries. On the relevance of theoretical syntax for the study of register variation. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication* 9(2). 157–200.

Haegeman, Liliane & Tabea Ihsane. 1999. Subject ellipsis in embedded clauses in English. *English Language and Linguistics* 3. 117–145

Haegeman, Liliane & Tabea Ihsane. 2001. Adult Null Subjects in the non-pro-drop Languages: Two Diary Dialects. *Language Acquisition* 9(4). 329–346.

Hagemeijer, Tjerk. 2011. The Gulf of Guinea creoles: genetic and typological relations. *Journal of Pidgin and Creole Languages* 26(1). 111–154.

Harris, Alice C. & Lyle Campbell. 1995. *Historical Syntax in Cross-Linguistic Perspective*. Cambridge University Press.

Harris, Catherine L. 2006. Language and Cognition. *Encyclopedia of Cognitive Science*. American Cancer Society.

Harvie, Dawn. 1998. Null subject in English: Wonder if it exists? *Cahiers Linguistique d'Ottawa* 26. 15–25.

Hashimoto, Daiki. 2019. Sociolinguistic effects on loanword phonology: Topic in speech and cultural image. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1).

Havik, Philip J. 2004. *Silences and soundbytes: the gendered dynamics of trade and brokerage in the precolonial Guinea Bissau region*. Münster: LIT.

Hawthorne, Walter. 2003. *Planting Rice and Harvesting Slaves: Transformations Along the Guinea-Bissau Coast, 1400-1900*. Heinemann. University of Michigan Press.

Heine, Bernd & Kyung-An Song. 2011. On the grammaticalization of personal pronouns. *Journal of Linguistics* 47(3). 587–630.

Heyselaar, Evelien, Katrien Segaert, Serge J.W. Walvoort, Roy P.C. Kessels & Peter Hagoort. 2017. The role of nondeclarative memory in the skill for language: Evidence from syntactic priming in patients with amnesia. *Neuropsychologia* 101.

Heywood, Linda M & John K Thornton. 2007. *Africans and Catholics: the first generation of African Americans in North America and the Caribbean, 1619-1660*. Cambridge: Cambridge University Press.

Hinrichs, Lars, Benedikt Szmrecsanyi & Axel Bohmann. 2015. Which-hunting and the Standard English relative clause. *Language* 91(4). 806–836.

Hogg, Richard (ed.). 1992. *The Cambridge History of the English Language*. Cambridge University Press.

Holler, Anke & Katja Suckow. 2016. Introduction. *Empirical Perspectives on Anaphora*

Resolution. Berlin: Walter de Gruyter.
Holmberg, Anders, Aarti Nayudu & Michelle Sheehan. 2009. Three partial null-subject languages: A comparison of Brazilian Portuguese, Finnish and Marathi. *Studia Linguistica* 63(1). 59–97.
Hopper, Paul J. & Elizabeth Closs Traugott. 2006. *Grammaticalization* (Cambridge Textbooks in Linguistics). 2. ed., 3. print. Cambridge: Cambridge Univ. Press.
Horta, José da Silva. 2000. Evidence for a Luso-African Identity in "Portuguese" Accounts on "Guinea of Cape Verde" (Sixteenth-Seventeenth Centuries). *History in Africa* 27. 99–130.
Hothorn, Torsten, Kurt Hornik & Achim Zeileis. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15(3). 651–674.
Huang, C.-T. James. 1984. On the Distribution and Reference of Empty Pronouns. *Linguistic Inquiry* 15(4). 531–574.
Huddleston, Rodney. 1988. *English grammar: An outline*. Cambridge University Press.
Hulk, Aafke, & Ans van Kemenade. 1995. Verb-second, pro-drop, functional projections, and langauge change. In Adrian Battye and Ian Roberts (ed.), *Clause structure and language change.* Oxford University Press.
Huttenlocher, Janellen, Marina Vasilyeva & Priya Shimpi. 2004. Syntactic priming in young children. *Journal of Memory and Language* 50(2). 182–195.
Jackendoff, Ray. 2015. The Parallel Architecture and its Place in Cognitive Science. In Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*, 645–668. 2nd edn. Oxford University Press.
Jacobs, Bart. 2010. Upper Guinea Creole: Evidence in favor of a Santiago birth. *Journal of Pidgin and Creole Languages* 25(2). 289–343.
Jacobs, Bart. 2012. *Origins of a Creole: The History of Papiamentu and Its African Ties*. Berlin: De Gruyter Mouton.
Jacobs, Bart & Nicolas Quint. 2016. On the relevance of Classical Portuguese features in four Atlantic creoles. In Armin Schwegler, John McWhorter & Liane Ströbel (eds.), *The Iberian Challenge*. Iberoamericana Vervuert.
Jaeggli, Osvaldo A. 1986. Passive. *Linguistic Inquiry* 17(4). 587–622.
Johnson, Keith. 2008. *Quantitative methods in linguistics*. Hoboken, NJ: Blackwell.
Jones, Katy Sarah. 2014. *Towards an understanding of the use of indefinite expressions for definite reference in English discourse*. Cardiff University. (Doctoral dissertation.)
Kaiser, Georg A. 2009. Losing the null subject. A contrastive study of (Brazilian) Portuguese and (Medieval) French. *Proceedings of the Workshop "Null-subjects, expletives, and locatives in Romance"* (Arbeitspapier), vol. 123, 26.
Kallen, Jeffrey L. 2005. Internal and external factors in phonological convergence: the case of English /t/ lenition. In Peter Auer, Frans Hinskens & Paul Kerswill (eds.), *Dialect Change: Convergence and Divergence in European Languages*. Cambridge University Press.
Kari, Ethelbert Emmanuel. 2017. On the Status of Subject Markers in African Languages. *Studies in Linguistics* null(44). 99–134.
Kassambara, Alboukadel. 2018. *Machine Learning Essentials: Practical Guide in R*. Marseilles: STHDA.
Kato, Mary Aizawa. 1999. Strong pronouns and weak pronominals in the null subject parameter. *Probus* 11(1). 1–37.
Kato, Mary Aizawa. 2000. The partial pro-drop nature and the restricted VS order in Brazilian

Portuguese. In Mary Aizawa Kato & Esmeralda Vailati Negrão (eds.), *Brazilian Portuguese and the Null Subject Parameter* (Ediciones de Iberoamericana 4). Frankfurt am Main/Madrid: Iberoamericana Vervuert.

Kato, Mary Aizawa & Maria Eugênia L. Duarte. 2003. Semantic and phonological constraints on the distribution of null subjects in Brazilian Portuguese. *32nd New Ways of Analyzing Variation.* Philadelphia.

Kato, Mary Aizawa & Maria Eugênia L. Duarte. 2005. (Micro)parametric variation between European Portuguese (EP) and Brazilian Portuguese (BP): Similarities and Differences Related to Ongoing Changes in Latin American Spanish. *Asociación de linguística y filología de América Latina.* Monterrey, MX.

Kato, Mary Aizawa & Maria Eugênia L. Duarte. 2013. The lack of lexical expletives in Brazilian Portuguese. *Encontro do GT de Teoria da Gramática*. São Paolo.

Kato, Mary Aizawa & Maria Eugênia L. Duarte. 2014. Restrições na distribuição de sujeitos nulos no Português brasileiro. *Web-Sociodialeto* 4(12): 153-177.

Kayne, Richard S. 1975. *French syntax: the transformational cycle* (Current Studies in Linguistics Series ; 6). Cambridge, Mass: MIT Press.

Kihm, Alain. 1989. Lexical Conflation as a Basis for Relexification. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 34(3). 351–376.

Kihm, Alain. 1994. *Kriyol Syntax* (Creole Language Library 14). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Klavans, Judith L. 1982. *Some Problems in Theory of Clitics*. Indiana University Linguistics Club.

Kouwenberg, Silvia. 1990. Complementizer PA, the Finiteness of its Complements and Some Remarks on Empty Categories in Papiamento. *Journal of Pidgin and Creole Languages* 5(1). 39–51.

Kouwenberg, Silvia. 1994. A Grammar of Berbice Dutch Creole. Berlin: Mouton de Gruyter.

Labov, William. 1972. The study of language in its social context. In William Labov (ed.), *Sociolinguistic Patterns*, 183–259. University of Pennsylvania Press.

Labov, William. 1978. Where Does the Linguistic Variable Stop? A Response to Beatriz Lavandera. *Working Papers in Sociolinguistics* 44.

Labov, William. 1982. Objectivity and Commitment in Linguistic Science: The Case of the Black English Trial in Ann Arbor. *Language in Society* 11(2). 165–201.

Labov, William. 1984. Field methods of the Project in Linguistic Change and Variation. In John Baugh & Joel Sherzer (eds.), *Language in use: readings in sociolinguistics*. Englewood Cliffs, NJ: Prentice-Hall.

Labov, William. 2004. Quantitative Reasoning in Linguistics. In U. Ammon, N. Dittmar, K.J. Mattheier & P. Trudgill (eds.), *Sociolinguistics/Soziolinguistik: An international handbook of the science of language and society*, vol. 1, 6–22. Berlin: Mouton de Gruyter.

Lambrecht, Knud. 1981. *Topic, Antitopic and Verb Agreement in Non-Standard French*. Amsterdam/Philadelphia: John Benjamins Publishing.

Lang, Jürgen. 1999. O pronome pessoal átono de primeira pessoa do singular e a nasalidade no crioulo de Santiago (Cabo Verde). In Klaus Zimmerman (ed.), *Las lenguas criollas de base lexical española y portuguesa*. Madrid/Frankfurt: Vervuert-Iberoamericana.

Lang, Jürgen. 2001. Breve esboço da gramática do crioulo da ilha de Santiago (Cabo Verde). *Santa Barbara Portuguese Studies* 5. 228–254.

Lang, Jürgen. 2003. Affinités grammaticales entre le créole portugais de l'Île de Santiago (Cap Vert) et le wolof (Sénegal) : aspect et temps. In Nancy Vázquez Veiga & Mauro Fernández

(eds.), *Los criollos de base ibérica ACBLPE 2003*, 137–147. Frankfurt: Iberoamericana Vervuert.
Lang, Jürgen. 2006. L'influence des Wolof et du wolof sur la formation du créole santiagais. In Jürgen Lang, John Holm & Jean-Louis Rougé (eds.), *Cabo Verde - origens da sua sociedade e do seu crioulo*. Tübingen: Gunter Narr Verlag.
Lang, Jürgen. 2009. *Les langues des autres dans la créolisation: théorie et exemplification par le créole d'empreinte wolof à l'île Santiago du Cap Vert*. Tübingen: Gunter Narr Verlag.
Lang, Jurgen. 2011. A Wolof trace in the verbal system of the Portuguese Creole of Santiago Island (Cape Verde). In Claire Lefebvre (ed.), *Creoles, their Substrates, and Language Typology* (Typological Studies in Langauge 95). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Lang, Jürgen. 2012. A filiação dos pronomes pessoais do crioulo da ilha de Santiago (Cabo Verde). *Journal of Ibero-Romance Creoles* 3.
Lavandera, Beatriz R. 1978. Where Does the Sociolinguistic Variable Stop? *Language in Society* 7(2). 171–182.
Lezama, Carlos Gelormini & Amit Almor. 2011. Repeated Names, Overt Pronouns, and Null Pronouns in Spanish. *Language and cognitive processes* 26(3). 437–454.
Li, Charles N. & Sandra A. Thompson. 1976. Subject and topic: A new typology of language. In Charles N. Li (ed.), *Subject and topic*, 457–489. New York: Academic Press.
Li, Xiaoshi, Xiaoqing Chen & Wen-Hsin Chen. 2012. Variation of subject pronominal expression in Mandarin Chinese. *Sociolinguistic Studies* 6(1).
Lipski, John. 1999. Null subjects in (Romance-derived) creoles: routes of evolution. *Meeting of the Society for Pidgin & Creole Languages.* Los Angeles, California.
Lobban, Richard Andrew. 1995. *Cape Verde: Crioulo colony to independent nation*. Westview Press.
Loebell, Helga & Kathryn Bock. 2003. Structural priming across languages. *Linguistics* 41(5).
Lopes, Carlos. 1999. *Kaabunké: espaço, território e poder na Guiné-Bissau, Gâmbia e Casamance pré-coloniais*. Lisboa: Comissão Nacional para as Comemorações dos Descobrimentos Portugueses.
Lopes da Silva, Baltasar. 1957. *O dialecto crioulo de Cabo Verde* (Junta Das Missões Geográficas e de Investigações Do Ultramar. Centro de Estudos Políticos e Sociais. Dialectosportugueses Do Ultramar). Lisboa: Imprensa Nacional de Lisboa.
Lüdecke, Daniel. 2019. *sjPlot: Data Visualization for Statistics in Social Science*. https://CRAN.R-project.org/package=sjPlot.
Luís, Ana R. & Georg A. Kaiser. 2016. Clitic Pronouns: Phonology, Morphology, and Syntax. In W. Leo Wetzels, João Costa & Sergio Menuzzi (eds.), *The Handbook of Portuguese Linguistics*, 210–233. 1st edn. Wiley.
Macqueen, Norrie. 1997. *The decolonization of Portuguese Africa: metropolitan revolution and the dissolution of empire*. Longman Publishing Group.
Maia, Clarinda Azevedo. 2017. *História do galego-português: estado linguístico da Galiza e do Noroeste de Portugal desde o século XIII ao século XVI: (com referência à situação do galego moderno)*. Coimbra: Imprensa da Universidade de Coimbra.
Mailer, Phil. 1977. *Portugal: the impossible revolution?* London: Solidarity.
Marins, Juliana. 2009. *O parâmetro do sujeto nulo: Uma análise contrastiva entre o português e o italiano*. Universidade Federal do Rio de Janeiro. (Doctoral dissertation.)
Mark, Peter. 1999. The Evolution of "Portuguese" Identity: Luso-Africans on the Upper Guinea

Coast from the Sixteenth to the Early Nineteenth Century. *The Journal of African History* 40(2). 173–191.
Mark, Peter. 2002. *Portuguese Style and Luso-African Identity*. Indiana University Press.
Martínez-Sanz, Cristina. 2011. *Null and overt subjects in a variable system: The case of Dominican Spanish*. University of Ottawa. (Doctoral dissertation.)
Martins, Ana Maria. 1994. *A história dos clíticos em português*. Universidade de Lisboa.
Maurer, Philippe. 1988. *Les modifications temporelles et modales du verbe dans le papiamento de Curaçao (Antilles néerlandaises): avec une anthologie et un vocabulaire papiamento-français* (Kreolische Bibliothek Bd. 9). Hamburg: H. Buske.
Mayer, Mercer. 1967. *A boy, a dog, and a frog*. Dial Press.
Mayol, Laia. 2010. Contrastive pronouns in null-subject Romance languages. *Lingua* 120(10). 2497–2514.
McWhorter, John H. 2005. *Defining Creole*. Oxford University Press.
Meijer, Paul J. A & Jean E Fox Tree. 2003. Building Syntactic Structures in Speaking: A Bilingual Exploration. *Experimental Psychology* 50(3). 184–195.
Meintel, Deirdre. 1984. *Race, culture, and Portuguese colonialism in Cabo Verde*. Maxwell School of Citizenship and Public Affairs, Syracuse University.
Milroy, Lesley. 1980. *Language and social networks*. London/Baltimore/Basil Blackwell: University Park Press.
Modesto, Marcello. 2000. Null Subjects without 'Rich' Agreement. In Mary Aizawa Kato & E Negrão (eds.), *Brazilian Portuguese and the Null Subject Parameter*, 147–174. Frankfurt: Vervuert-Iberoamericana.
Modesto, Marcello. 2008. Topic prominence and null subjects. In Theresa Biberauer (ed.), *Linguistik Aktuell/Linguistics Today*, vol. 132, 375–409. Amsterdam: John Benjamins Publishing Company.
Montalbetti, Mario M. 1984. *After binding : on the interpretation of pronouns*. Massachusetts Institute of Technology. (Doctoral dissertation.)
Mota, Avelino Teixeira da. 1954. *Guiné portuguesa, pelo primeiro tenente A. Teixeira da Mota* Lisboa: Agencia geral do Ultramar. Divisão de publicações e biblioteca.
Mota, Avelino Teixeira da. 1977. *Descrição da Serra Leoa e dos rios de Guiné do Cabo Verde,1625 =: An account of Sierra Leone and the rivers of Guinea of Cape Verde, 1625* ([Série de memórias] - Centro de Estudos de Cartografia Antiga, Secção de Lisboa ; 19). (Ed.) P.E.H. Hair. Lisboa: Junta de Investigações Científicas do Ultramar.
Mufwene, Salikoko S. 1996. The Founder Principle in Creole Genesis. *Diachronica* 13(1). 83–134.
Mufwene, Salikoko S. 1998. The Small pro and Inflectional Morphology. *Journal of Pidgin and Creole Languages* 13(1). 213–213.
Mufwene, Salikoko S. & Marta Dijkhoff. 1989. On the so-called 'infinitive' in Atlantic creoles. *Lingua* 77(3–4).
Mufwene, Salikoko S. 2001. *The Ecology of Language Evolution*. Cambridge University Press.
Munslow, Barry. 1981. The 1980 Coup in Guinea Bissau. *Review of African Political Economy* 21. 109-113.
Muysken, Pieter. 2001. Creolization. *Language Typology and Language Universals: an international handbook*, 1656–1667. Berlin: de Gruyter.
Muysken, Pieter. 2008. *Functional Categories*. Cambridge University Press.
Nadasdi, Terry. 1995. *Variation morphosyntaxique et langue minoritaire: le cas du français*

ontarien. University of Toronto. (Doctoral dissertation.)
Nafafe, José Lingna. 2007. Lançados, Culture and Identity: Prelude to Creole Societies on the Rivers of Guinea and Cape Verde. In Philip J Havik & Malyn Newitt (eds.), *Creole Societies in the Portuguese Colonial Empire.*, 65–91. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.
Naimi, Babak, Nicholas A. S. Hamm, Thomas A. Groen, Andrew K. Skidmore & Albertus G. Toxopeus. 2014. Where is positional uncertainty a problem for species distribution modelling? *Ecography* 37(2). 191–203.
Namiuti, Cristiane. 2008. *Aspectos da História Gramatical do Português: Interpolação, Negação e Mudança*. Universidade Estadual de Campinas.
Naro, Anthony J. 1978. A Study on the Origins of Pidginization. *Language* 54(2). 314–347.
Negrão, Esmeralda Vailati & Evani Viotti. 2000. Brazilian Portuguese as a Discourse-Oriented Language. In Mary Aizawa Kato & Esmeralda Vailati Negrão (eds.), *Brazilian Portuguese and the null subject parameter* (Sprachwissenschaft/Lingüística 4). Madrid/Frankfurt: Ediciones de Iberoamericana.
Neves, Ana Cristina. 2009. Younger Learners and Phonological Variation in the Portuguese Mesolect of Cape Verde. *Linguistica Online* 14.
Newitt, Malyn. 2004. *A History of Portuguese Overseas Expansion 1400–1668*. Routledge.
Newitt, Malyn. 2015. *Emigration and the Sea: An Alternative History of Portugal and the Portuguese*. Oxford University Press.
Nicolis, Marco. 2008. The null subject parameter and correlating properties: The case of Creole languages. In Theresa Biberauer (ed.), *The Limits of Syntactic Variation* (Linguistik Aktuell/Linguistics Today 132), 271–294. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Nikolaeva, Irina (ed.). 2007. *Finiteness: theoretical and empirical foundations* (Oxford Linguistics). Oxford ; New York: Oxford University Press.
Olson, Clif. 1992. Gumawana (Amphlett Islands, Papua New Guinea): grammar sketch and texts. *Papers in Austronesian Linguistics*, vol. 2, 251–430. Canberra: Australian National University.
O'Malley, Gregory E. & Alex Borucki. 2017. Patterns in the intercolonial slave trade across the Americas before the nineteenth century. *Tempo* 23(2). 314–338.
Onishi, Kristine H, Kyle E Chambers & Cynthia Fisher. 2002. Learning phonotactic constraints from brief auditory experience. *Cognition* 83(1). B13–B23.
Oquendo, Melissa. 2014. The Sociolinguistic Situation and National Identity of Cape Verde. *Bridgewater State University Undergraduate Review* (Anthropological Linguistics and Sociolinguistics Commons) 10. 116–119.
Ordóñez, Francisco & Antonio- Olarrea. 2006. Microvariation in Caribbean/non Caribbean Spanish interrogatives. *Probus* 18(1). 59–96.
Otheguy, Ricardo & Ana Celia Zentella. 2012. *Spanish in New York: language contact, dialectal leveling, and structural continuity* (Oxford Studies in Sociolinguistics). Oxford ; New York: Oxford University Press.
Otheguy, Ricardo, Ana Celia Zentella & David Livert. 2007. Language and Dialect Contact in Spanish in New York: Toward the Formation of a Speech Community. *Language* 83(4). 770–802.
Paixão de Sousa, Maria Clara. 2004. *Lingua Barroca : sintaxe e historia do portugues nos 1600*. Universidade Estadual de Campinas. (Doctoral dissertation.)

Palasis, Katerina. 2015. Subject clitics and preverbal negation in European French: Variation, acquisition, diatopy and diachrony. *Lingua* 161. 125–143.
Parkvall, Mikael. 2000. *Out of Africa: African Influences in Atlantic Creoles*. Battlebridge Publications.
Passos, Joana Filipa da Silva de Melo Vilela. 2003. *Micro-universes and Situated Critical Theory: Postcolonial and Feminist Dialogues in a Comparative Study of Indo-English and Lusophone Women Writers*. Utrecht: Proefschrift Universiteit.
Patterson, K. David. 1988. Epidemics, Famines, and Population in the Cape Verde Islands, 1580-1900. *The International Journal of African Historical Studies* 21(2). 291.
Peck, Stephen Madry. 1988. *Tense, aspect and mood in Guinea-Casamance Portuguese Creole*. University of California, Los Angeles.
Pereira, Dulce. 2006. Contributos da História Geral de Cabo Verde para o estudo da formção e da difusão do crioulo caboverdiano. In Jürgen Lang, John Holm & Jean-Louis Rougé (eds.), *Cabo Verde - origens da sua sociedade e do seu crioulo*. Tübingen: Gunter Narr Verlag.
Perlmutter, David M. 1971. *Deep and surface structure constraints in syntax* (The Transatlantic Series in Linguistics). New York: Holt, Rinehart and Winston.
Pescarini, Diego. 2018. Pronoun systems across Romance. *Oxford Encyclopedia of Romance Linguistics*, 21. Oxford University Press.
Pickering, Martin J. & Victor S. Ferreira. 2008. Structural Priming: A Critical Review. *Psychological bulletin* 134(3). 427–459. doi:10.1037/0033-2909.134.3.427.
Pituch, Keenan A. & James Stevens. 2016. *Applied multivariate statistics for the social sciences: analyses with SAS and IBM's SPSS*. 6th edition. New York: Routledge/Taylor & Francis Group.
Poletto, Cecilia. 2000. *The higher functional field: evidence from northern Italian dialects* (Oxford Studies in Comparative Syntax). New York: Oxford University Press.
Poletto, Cecilia & Christina Tortora. 2016. Subject clitics: syntax. In Adam Ledgeway & Martin Maiden (eds.), *The Oxford Guide to the Romance Languages*, 772–785. Oxford University Press.
Pontes, Eunice. 1987. *O tópico no português do Brasil*. Campinas: Pontes.
Poplack, Shana. 1980. Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching. *Linguistics* 18(7–8).
Poplack, Shana. 1993. Variation theory and language contact: Concepts, methods and data. In Dennis R. Preston (ed.), *American Dialect Research: Celebrating the 100th anniversary of the American Dialect Society, 1889–1989*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Posio, Pekka. 2014. Subject expression in grammaticalizing constructions: The case of creo and acho 'I think' in Spanish and Portuguese. *Journal of Pragmatics* 63. 5–18.
Powell, M. J. D. 2009. *The BOBYQA algorithm for bound constrained optimization without derivatives*. Department of Applied Mathematics and Theoretical Physics, Cambridge University.
Prada Pérez, Ana de. 2009. *Subject Expression in Minorcan Spanish: Consequences of Contact with Catalan*. The Pennsylvania State University. (Doctoral dissertation.)
Pratas, Fernanda. 2004. *O sistema pronominal do caboverdiano (variante de Santiago) : questões de gramática*. Lisboa: Edições Colibri.
Quint, Nicolas. 2000a. *Grammaire de la langue cap-verdienne: étude descriptive et compréhensive du créole afro-portugais des Iles du Cap-Vert*. Paris: L'Harmattan.

Quint, Nicolas. 2000b. *Le cap-verdien: origines et devenir d'une langue métisse: étude des relations de la langue cap-verdienne avec les langues africaines, créoles et portugaise*. Paris, France: L'Harmattan.

Quint, Nicolas. 2006. Un bref aperçu des racines africaines de la langue capverdienne. In Jürgen Lang, John Holm & Jean-Louis Rougé (eds.), *Cabo Verde - origens da sua sociedade e do seu crioulo*. Tübingen: Gunter Narr Verlag.

Quint, Nicolas. 2008a. *L'élément africain dans la langue capverdienne. Bilingue portugais-français*. L'Harmattan.

Quint, Nicolas. 2008b. A realização do sujeito em português do Brasil : deriva versus crioulização. in J.L. Fiorin and M. Petter, (eds.), *África no Brasil*. São Paulo: Contexto. 75-88.

Quint, Nicolas. 2008c. Les Apontamentos de António de Paula Brito (1887) ou la naissance d'une tradition grammaticale capverdienne autochtone. *Histoire Épistémologie Langage* 30(1). 127–153.

Quint, Nicolas. 2009. O cabo-verdiano : uma língua mundial. *Revista de Estudos Caboverdianos* (3). 129–144.

Quint, Nicolas. 2012. Les influences du portugais contemporain sur le système verbal du capverdien santiagais. In Chamreau Claudine & Goury Laurence (eds.), *Changement linguistique et langues en contact : approches plurielles du domaine prédicatif*, 155–178. CNRS.

Quint, Nicolas. 2015. *Let's Speak Capeverdean: Language and Culture*. Battlebridge Publications.

R Core Team. 2019. *The R Project for Statistical Computing*.

Ramos, Jânia. 1997. O uso das formas você, ocê e cê no dialeto mineiro. In Dermeval da Hora (ed.), *Diversidade lingüística no Brasil.*, 43–60. João Pessoa: Idéia.

Ramos Moreira, Cesário. 2005. *Geologia económica do Concelho de Santa Cruz*. Lisbon, Portugal: Instituto Superior de Educação e Ciências (Bachelors of Science thesis.)

Rao, Rajiv & Sandro Sessarego. 2016. On the Intonation of Afro-Bolivian Spanish Declaratives: Implications for a Theory of Afro-Hispanic Creole Genesis. *Lingua* 174. 45–64.

Rao, Rajiv & Sandro Sessarego. 2018. Intonation in Chota Valley Spanish: Contact-Induced Phenomena at the Discourse/Phonology Interface. *Studies in Hispanic and Lusophone Linguistics* 11(1). 163–192.

Rego, Márcia. 2015. *The Dialogic Nation of Cape Verde: Slavery, Language, and Ideology*. Lexington Books.

Renzi, Lorenzo & Laura Vanelli. 1983. I pronomi soggetto in alcune varietà romanze. *Scritti Linguistici in onore di G.B. Pellegrini*, 121–145. Piacini, Piaza.

Ribeiro, Ilza. 1995. *Clause Structure and Language Change* (Generative Diachronic Syntax Conference 1990 (University Of). (Ed.) Adrian Battye & Ian Roberts. Oxford University Press.

Rickford, J. & F. McNair-Knox. 1994. Addressee- and topic-influenced style shift: A quantitative sociolinguistic study. In Douglas Biber & Edward Finegan (eds.), *Sociolinguistic perspectives on register*, vol. Oxford studies in sociolinguistics. New York: Oxford University Press.

Rickford, John R. 1975. Carrying the New Wave into Syntax: The Case of Black English Been. In Ralph W. Fasold (ed.), *Variation in the Form and Use of Language.*, 98–119. Washington, D.C: Georgetown University Press.

Riiho, Timo. 1994. Evolução Linguística Interna. In Michael Metzeltin & Christian Schmitt (eds.),

Lexikon der Romanistischen Linguistik (LRL) VI, vol. 2, 498–511. Tübingen: Niemeyer.
Rivero, María-Luisa. 1993. Long head movement vs. V2, and null subjects in old Romance. *Lingua* (Special Issue Null Subjects in Diachrony) 89(2). 217–245.
Rizzi, Luigi. 1982. *Issues in Italian Syntax*. De Gruyter.
Rizzi, Luigi. 1986a. On the status of subject clitics in romance. In Osvaldo A. Jaeggli & Carmen Silva-Corvalán (eds.), *Studies in Romance linguistics* (Publications in Language Sciences 24), 391–419. Dordrecht: Foris.
Rizzi, Luigi. 1986b. Null Objects in Italian and the Theory of pro. *Linguistic Inquiry* 17(3). 501–557.
Rizzi, Luigi. 1997. The Fine Structure of the Left Periphery. In Liliane Haegeman (ed.), *Elements of Grammar: Handbook in Generative Syntax* (Kluwer International Handbooks of Linguistics), 281–337. Dordrecht: Springer Netherlands.
Roberge, Yves. 1990. *Syntactic Recoverability of Null Arguments*. McGill-Queen's Press - MQUP.
Roberts, Ian G. 1993. *Verbs and Diachronic Syntax: A Comparative History of English and French* (Studies in Natural Language and Linguistic Theory 28). Dordrecht: Springer.
Roberts, Ian & Anders Holmberg. 2009. Introduction: parameters in minimalist theory. In Theresa Biberauer, Anders Holmberg, Ian Roberts & Michelle Sheehan (eds.), *Parametric Variation: Null subjects in Minimalist Theory*. Cambridge University Press.
Rodney, Walter. 1970. *A History of the Upper Guinea Coast, 1545-1800*. Monthly Review Press.
Romaine, Suzanne. 1980. Stylistic Variation and Evaluative Reactions To Speech: Problems in the Investigation of Linguistic Attitudes in Scotland.
Rosemeyer, Malte & Andrés Enrique-Arias. 2016. A match made in heaven: Using parallel corpora and multinomial logistic regression to analyze the expression of possession in Old Spanish. *Language Variation and Change* 28(3). 307–334.
Rosemeyer, Malte & Scott A. Schwenter. 2019. Echoic and non-echoic confirming affirmative responses in spoken Brazilian Portuguese. *Journal of Pragmatics* 141. 80–101.
Rothman, Jason & Roumyana Slabakova. 2011. The Mind-Context Divide: On acquisition at the linguistic interfaces. *Lingua* (Acquisition at the Linguistic Interfaces) 121(4). 568–576.
Rougé, Jean-Louis. 1993. Crioulos. *Littérature du Cap-Vert, de Guinée-Bissau, de São Tomé et de Principe* (special edition of *Notre Librairie*), 112–113.
Rougé, Jean-Louise. 1994. A propos de la formation des créoles du Cap Vert et de Guinée. *Papia* 3(2). 137–149.
Rougé, Jean-Louis. 1999. Apontamentos sobre o léxico de origem africana dos crioulos da Guiné e de Cabo Verde (Santiago). *Lenguas criollas de base lexical espanola y portuguesa*, 49–66. Colloquium Verlag.
Rougé, Jean-Louis. 2006. L'influence manding sur la formation des créoles du Cap-Vert et de Guinée-Bissau et Cassamance. In Jürgen Lang, John Holm & Jean-Louis Rougé (eds.), *Cabo Verde - origens da sua sociedade e do seu crioulo*. Tübingen: Gunter Narr Verlag.
Rowlands, Evan Colyn. 1959. *A grammar of Gambian Mandinka*. London: University of London.
Salvi, Giampolo. 2000. La Formazione Della Sistema V2 Delle Lingue Romanze Antiche. *Lingua E Stile* 35(4). 665–692.
Sánchez, Liliana. 2003. *Quechua-Spanish Bilingualism*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Sankoff, David. & Stephen Laberge. 1978. The linguistic market and the statistical explanation of variability. In David Sankoff (ed.), *Linguistic variation: models and methods*, 239–250. New York: Academic Press.

Sankoff, David. 1973. Above and beyond phonology in variable rules. In Southeastern Conference on Linguistics, Charles James Nice Bailey & Roger W. Shuy (eds.), *New ways of analyzing variation in English*. Washington: Georgetown University Press.

Sankoff, David. 1988. Problems of representativeness. In Ulrich Ammon, Norbert Dittmar & Klaus Mattheier (eds.), *Sociolinguistics: an international handbook of the science of language and society = Soziolinguistik: ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft* (Handbücher Zur Sprach- Und Kommunikationswissenschaft ; Handbooks of Linguistics and Communication Science = Manuels de Linguistique et Des Sciences de Communication Bd. 3.1). Berlin ; New York: W. de Gruyter.

Savoia, Leonardo M & M Rita Manzini. 2010. Les clitiques sujets dans les variétés occitanes et francoprovençales italiennes. *Corpus Linguistics and Linguistic Theory* 9. 27.

Schilling-Estes, Natalie. 2004. Constructing ethnicity in interaction. *Journal of Sociolinguistics* 8(2). 163–195.

Scott, Candice L. 2016. *Tense & Aspect Markers in African American English.* The University of Michigan. (Doctoral dissertation.)

Seibert, Gerhard. 2014. Crioulização em Cabo Verde e São Tomé e Príncipe: divergências históricas e identitárias. *Afro-Ásia* (49). 41–70.

Sessarego, Sandro. 2011. On the Status of Afro-Bolivian Spanish Features: Decreolization or Vernacular Universals? In Jim Michnowicz (ed.), *Proceedings of The 5th International Workshop on Spanish Sociolinguistics (WSS5)*, 125–141. Somerville, MA: Cascadilla.

Sessarego, Sandro. 2012. The Contribution of Afro-Hispanic Contact Varieties to the Study of Syntactic Microvariation. In Melvin González-Rivera & Sandro Sessarego (eds.), *Current Formal Aspects of Spanish Syntax and Semantics*, 229–250. Newcastle: Cambridge Scholars Publishing.

Sessarego, Sandro. 2013a. Afro-Hispanic contact varieties as conventionalized advanced second languages. *IBERIA: An International Journal of Theoretical Linguistics* 5(1). 99–125.

Sessarego, Sandro. 2013b. Enhancing Dialogue between Quantitative Sociolinguistics and Minimalist Syntax. *Studies in Hispanic and Lusophone Linguistics* 5(2). 79–97.

Sessarego, Sandro. 2014a. Afro-Peruvian Spanish in the context of Spanish Creole Genesis. *Spanish in Context* 11(3). 381–401.

Sessarego, Sandro. 2014b. *The Afro-Bolivian Spanish Determiner Phrase: A Microparametric Account*. Ohio State University Press.

Sessarego, Sandro. 2017. A Feature-Geometry Account for Subject-Verb Agreement Phenomena in Yungueño Spanish. In J Colomina-Almiñana (ed.), *Contemporary Studies on Theoretical and Applied Linguistics of Spanish Variation*, 265–282. Columbus, OH: Ohio State University Press.

Sessarego, Sandro. 2019. Universal processes in contact-induced syntactic change. *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK)*, vol. 45. 1st edn. Berlin/Boston: de Gruyter Mouton.

Sessarego, Sandro & L Ferreira. 2016. Spanish and Portuguese Parallels: Impoverished Number Agreement as a Vernacular Feature of Two Rural Dialects. In Sandro Sessarego & F Tejedo-Herrero (eds.), *Spanish Language and Sociolinguistic Analysis*, 283–304. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sessarego, Sandro & Javier Gutiérrez-Rexach. 2010. Variation, evolution and the syntax of Afro-Bolivian Spanish gender agreement. *Revista de Crioulos de Base Lexical Portuguesa e*

Espanhola 2. 63–86.
Sessarego, Sandro & Javier Gutiérrez-Rexach. 2011. A Minimalist approach to gender agreement in the Afro-Bolivian DP: Variation and the specification of uninterpretable features. *Folia Linguistica* 45(2). 465–488.
Sessarego, Sandro & Javier Gutiérrez-Rexach. 2017. Revisiting the Null Subject Parameter: New Insights from Afro-Peruvian Spanish. *Isogloss. A journal on variation of Romance and Iberian languages* 3(1). 43–68.
Sessarego, Sandro & Javier Gutiérrez-Rexach. 2018. Afro-Hispanic Contact Varieties at the Interface: A Closer Look at Pro-Drop Phenomena in Chinchano Spanish. In Jeremy King & Sandro Sessarego (eds.), *Language Variation and Contact-Induced Change: The Spanish Language across Space and Time*, 85–110. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Sessarego, Sandro & Rajiv Rao. 2016. On the Simplification of a Prosodic Inventory: The Afro-Bolivian Spanish Case. In Alejandro Cuza, L Czerwionka & Dan Olson (eds.), *Inquiries in Hispanic Linguistics*, 171–190. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Sessarego, Sandro & Adrián Rodríguez-Riccelli. 2018. Formal issues in Afro-Hispanic morphosyntax: The Afro-Bolivian Spanish case. *Lingua* 202. 58–75.
Siewierska, Anna. 1999. From anaphoric pronoun to grammatical agreement marker: Why objects don't make it. *Folia Linguistica* 33(1–2).
Siewierska, Anna. 2004. *Person* (Cambridge Textbooks in Linguistics). Cambridge ; New York: Cambridge University Press.
Silva, Andressa Christine Oliveira da & Aline Alves Fonseca. 2018. Prosódia e processamento: compreensão e produção de estruturas de tópico e de sujeito no português brasileiro. *Revista de Estudos da Linguagem* 26(4). 1601–1646.
Silva, Claudia Roberta Tavares, Danniel Da Silva Carvalho & Fernanda Maciel Ziober. 2018. Licenciamento de duplos sujeitos em variedades do português: pessoa, definitude e estrutura de traços. *Letras Escreve* 7(2). 91.
Silva, Gláucia V. 2013. *Word Order in Brazilian Portuguese*. Reprint 2012. Berlin, Boston: De Gruyter Mouton.
Silva, Izione Santos. 1985. *Variation and change in the verbal system of Capeverdean Crioulo*. Washington, D.C.: Georgetown University.
Silva, Izione Santos. 1990. Tense and Aspect in Capeverdean Crioulo. In John Victor Singler (ed.), *Pidgin and Creole Tense/Mood/Aspect Systems* (Creole Language Library 6). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Silva, Rosa Virgínia Mattos (ed.). 2008a. *O Português arcaico: uma aproximação - Léxico e morfologia*. Vol. 1. Lisbo: Imprensa Nacional-Casa da Moeda.
Silva, Rosa Virgínia Mattos e (ed.). 2008b. *O Português arcaico: uma aproximação - Sintaxe e fonologia*. Vol. 2. Lisboa: Imprensa Nacional-Casa da Moeda.
Singler, John Victor. 2009. The Sociohistorical Context of Creole Genesis. *The Handbook of Pidgin and Creole Studies*, 332–358. John Wiley & Sons, Ltd.
Sippola, Eeva. 2011. *Una gramática descriptiva del chabacano de Ternate - Forskning - Aarhus Universitet*. University of Helsinki. (Docotral dissertation.)
Sitaridou, Ioanna. 2012. A comparative study of word order in Old Romance. In A Carlier, B Lamiroy & W. de Mulder (eds.), *Folia Linguistica: The pace of Grammaticalisation in Romance*, vol. 45, 553–604. 2nd edn.

Slabakova, Roumyana. 2006. Is there a critical period for semantics? *Second Language Research* 22(3). 302–338.
Slabakova, Roumyana. 2009. What Is Easy and What Is Hard to Acquire in a Second Language? In Melissa Bowles, Tania Ionin, Silvina Montrul & Annie Tremblay (eds.), *Proceedings of the 10th Generative Approaches to Second Language AcquisitionConference (GASLA 2009)*, 16. Somerville, MA: Cascadilla Proceedings Project.
Slobin, Dan I. 2004. The Many Ways to Search for a Frog: Linguistic Typology and the Expression of Motion Events. *Relating events in narrative, Volume 2: Typological and contextual perspectives*, 219–257. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
Soares da Silva, Humberto. 2006. *O parâmetro do sujeito nulo: confronto entre o português e o espanhol*. Universidade Federal do Rio de Janeiro.
Sodacı, Hande, Ad Backus & Gerrit Jan Kootstra. 2019. *Role of structural priming in contact-induced change: Subject pronoun expression in L1 Turkish by Turkish-Dutch bilinguals*. Preprint. PsyArXiv.
Sorace, Antonella. 2003. Near-Nativeness. *The Handbook of Second Language Acquisition*, 130–151. John Wiley & Sons, Ltd.
Sorace, Antonella. 2011. Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism* 1. 1–33.
Sorace, Antonella, Ludovica Serratrice, Francesca Filiaci & Michela Baldo. 2009. Discourse conditions on subject pronoun realization: Testing the linguistic intuitions of older bilingual children. *Lingua* 119(3). 460–477.
Speyer, Augustin. 2016. Factors determining the choice of anaphora in Old High German – A survey of zero and personal pronoun usage in Otfrid. In Anke Holler & Katja Suckow (eds.), *Empirical Perspectives on Anaphora Resolution*, vol. 563. Berlin, Boston: De Gruyter.
Spina, Segismundo (ed.). 2017. *História da Língua Portuguesa*. 1st edn. Ateliê Editorial.
Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin & Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9(1). 307.
Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis & Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8(1). 25.
Strömqvist, Sven, Ludo Verhoeven & Dan I. Slobin (eds.). 2004. The many ways to search for a frog: Linguistic typology and the expression ofmotion events. *Relating Events in Narrative, Volume 2: Typological and Contextual Perspectives*, 219–257. Mahwah, NJ, US: Taylor & Francis.
Suñer, Margarita. 1983. Free relatives and the pro head hypothesis. In W Harbet (ed.), *Cornell Working Papers*, vol. 4, 223–248.
Syea, Anand. 1993. Null Subject in Mauritian Creole and the Pro-Drop Parameter. In Francis Byrne & John Holm (eds.), *Atlantic Meets Pacific: A global view of pidginization and creolization* (Creole Language Library 11). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Szmrecsanyi, Benedikt. 2005. Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1(1).
Szmrecsanyi, Benedikt. 2006. *Morphosyntactic Persistence in Spoken English: A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis* (Trends in Linguistics. Studies and Monographs [TiLSM]). Berlin, New York: Mouton de

Gruyter.
Tagliamonte, Sali. 1998. Was/were variation across the generations: View from the city of York. *Language Variation and Change* 10(2). 153–191. doi:10.1017/S0954394500001277.
Tagliamonte, Sali A. 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.
Tagliamonte, Sali. 2012. *Variationist Sociolinguistics: Change, Observation, Interpretation | Wiley*. Wiley-Blackwell.
Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.
Tarallo, Fernando. 2015. Por uma sociolingüística românica "paramétrica": fonologia e sintaxe. *Cadernos de Linguística e Teoria da Literatura* 7(13). 53–85.
Tarallo, Fernando & Mary Kato. 2007. Harmonia trans-sistêmica: variação intra- e inter-linguística. *Revista Diadorim* 2.
Tavares, Bernadino. 2012. *The Verbal System of the Cape Verdean Creole of Tarrafal, Santiago: A Semantic Analysis of the Tense, Mood and Aspect Markers*. Faculdade de Letras, Universidade de Coimbra. (Docotral dissertation.)
Tavares Silva, Claudia Roberta, Danniel da Silva Carvalho & Fernanda Maciel Ziober. 2018. Licienciamento de duplos sujeitos em variedades do português: pessoa definitude, e estrutura de traços. *Letras Escreve* 7(2). 91.
Teixeira, Eliana Pitombo. 2012. A representação do sujeito pronominal no português popular angolano // Representation of the pronominal subject in vernacular Angolan Portuguese. *PAPIA - Revista Brasileira de Estudos do Contato Linguístico* 22(1). 141–159.
Teyssier, Paul. 1959. *La langue de Gil Vicente*. Paris: C. Klincksieck.
Teyssier, Paul. 1980. *Histoire de la langue portugaise*. Presses universitaires de France.
Thomas, Hugh. 1997. *The Slave Trade: The Story of the Atlantic Slave Trade 1440-1870*. New York: Simon & Schuster.
Thornton, John K. 1998. *Africa and Africans in the making of the Atlantic world, 1400-1800*. Cambridge University Press.
Toribio, Almeida Jacqueline. 1993. *Parametric variation in the licensing of nominals*. Cornell University. (Doctoral dissertation.)
Toribio, Almeida Jacqueline. 2000. Setting parametric limits on dialectal variation in Spanish. *Lingua* 110(5). 315–341.
Torrão, Maria Manuel. 2007. Rotas cemerciais, agentes económicos, meios de pagamento. In Maria Emília Madeira Santos, Maria Manuel Torrão & Maria João Soares (eds.), *História concisa de Cabo Verde*, vol. 2. 2nd edn. Instituto de Investigação Científica Tropical.
Torrence, Harold. 2013a. *The clause structure of Wolof: insights into the left periphery* (Linguistik Aktuell/Linguistics Today 198). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Torrence, Harold. 2013b. A Promotion Analysis of Wolof Clefts. *Syntax* 16(2). 176–215.
Torrence, William Harold. 2005. *On the distribution of complementizers in Wolof*. University of California, Los Angeles. (Doctoral dissertation.)
Torres Cacoullos, Rena & Catherine E. Travis. 2015. Foundations for the study of subject pronoun expression in Spanish in contact with English: Assessing inter-linguistic (dis)similarity via intra-linguistic variability. In Ana M. Carvalho, Rafael Orozco & Naomi Lapidus Shin (eds.), *Subject pronoun expression in Spanish*, 83-102. Georgetown University Press.

Torres Cacoullos, Rena & Catherine E. Travis. 2014. Prosody, priming and particular constructions: The patterning of English first-person singular subject expression in conversation. *Journal of Pragmatics* (Discourse Participants in Interaction: Cross-Linguistic Perspectives on Subject Expression and Ellipsis) 63. 19–34.
Torres Cacoullos, Rena & Catherine E. Travis. 2016. Two languages, one effect: Structural priming in spontaneous code-switching. *Bilingualism: Language and Cognition* 19(4). 733–753.
Torres Cacoullos, Rena & Catherine E. Travis. 2018. *Bilingualism in the Community: Code-switching and Grammars in Contact*. Cambridge University Press.
Torres Cacoullos, Rena & Catherine E. Travis. 2019. Variationist typology: Shared probabilistic constraints across (non-)null subject languages. *Linguistics* 57(3). 653–692.
Torres Moraes, Maria Aparecida. 1995. *Do português clássico ao português moderno: um estudo da cliticização e do movimento do verbo*. Universidade Estadual de Campinas. (Doctoral dissertation.)
Travis, Catherine E. 2007. Genre effects on subject expression in Spanish: Priming in narrative and conversation. *Language Variation and Change* 19(2). 101–135.
Travis, Catherine E. & Amy M. Lindstrom. 2016. Different registers, different grammars? Subject expression in English conversation and narrative. *Language Variation and Change* 28(1).
Travis, Catherine E. & Rena Torres Cacoullos. 2018. Discovering structure: Person and accessibility. In Naomi L. Shin & Daniel Erker (eds.), *Studies in Functional and Structural Linguistics*, vol. 76, 67–90. Amsterdam/Philadelphia: John Benjamins Publishing Company.
Tsimpli, Lanthi & Antonella Sorace. 2006. Differentiating interfaces: L2 performance in syntax-semantics and syntax-discourse phenomena. In David Bamman, Tatiana Magnitskaia & Colleen Zaller (eds.), *Proceedings of the 30th annual Boston University Conference on Language Development*, 653–664. Cascadilla Proceedings Project.
Tuck, Michael W. 2012. Everyday Commodities, the Rivers of Guinea, and the Atlantic World: The Beeswax Export Trade, c.1450–c.1800. *Brokers of Change: Atlantic Commerce and Cultures in Pre-Colonial Western Africa*. British Academy.
UNESCO Institute of Statistics. 2016. *Cabo Verde*.
UNICEF. 2017. *UNICEF Annual Report 2017: Cabo Verde*. UNICEF.
Vainikka, Anne & Yonata Levy. 1999. Empty Subjects in Finnish and Hebrew. *Natural Language & Linguistic Theory* 17(3). 613–671.
Vance, Barbara S. 1997. *Syntactic change in medieval French verb-second and null subjects* (Studies in Natural Language and Linguistic Theory v. 41). Dordrecht ; Boston: Kluwer Academic Publishers.
Veenstra, Tonjes. 1996. *Serial verbs in Saramaccan: Predication and creole genesis*. Dordrecht: ICG Printing.
Veiga, Manuel. 1996. *O crioulo de Cabo Verde: introdução à gramática*. Istituto Caboverdeano do Livro e do Disco.
Veiga, Manuel. 2000. *Le créole du Cap-Vert : étude grammaticale descriptive et contrastive*. Paris: Karthala.
Veiga, Manuel. 2002. *O Caboverdiano em 45 lições: estudo sociolinguístico e gramatical*. Praia, Cape Verde: INIC.
Veiga, Manuel. 2004. *A construção do bilinguismo*. Instituto da Biblioteca Nacional e do Livro.
Venables, W. N. & B. D. Ripley. 2002. *Modern Applied Statistics with S*. New York: Springer.

Ventura, Maria da Graça Mateus. 1999. *Negreiros portugueses na rota das Índias de Castela: 1541-1556*. Lisboa: Colibrí: Instituto de Cultura Ibero-Atlântica.
Verdu, Paul, Ethan M. Jewett, Trevor J. Pemberton, Noah A. Rosenberg & Marlyse Baptista. 2017. Parallel Trajectories of Genetic and Linguistic Admixture in a Genetically Admixed Creole Population. *Current Biology* 27(16). 2529–2535.
Villa-García, Julio & Imanol Suárez-Palma. 2016. Early null and overt subjects in the Spanish of simultaneous English-Spanish bilinguals and Crosslinguistic Influence. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics* 29(2). 350–395.
Vitral, Lorenzo. 1996. A forma CÊ e a noção de gramaticalização. *Revista de Estudos da Linguagem* 4(1). 115–124.
Wagner, Susanne. 2012. *Null subjects in English: Variable rules, variable language?* Institut für Anglistik und Amerikanistik, Chemnitz University of Technology Postdoctoral dissertation (Habilitationsschrift).
Wagner, Susanne. 2016. Never saw one – first-person null subjects in spoken English. *English Language and Linguistics* 22(01). 1–34. doi:10.1017/S1360674316000216.
Weiner, E. Judith & William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19(1). 29–58.
Weinreich, Uriel, William Labov & Marvin Herzog. 1968. *Empirical foundations for a theory of language change*. Austin: University of Texas Press.
Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag.
Wilson, John. 1987. The sociolinguistic paradox: Data as a methodological product. *Language & Communication* 7(2). 161–177.
Wilson, John & Alison Henry. 1998. Parameter setting within a socially realistic linguistics. *Language in Society* 27(1). 1–21.
Winford, Donald. 1984. The linguistic variable and syntactic variation in creole continua. *Lingua* 62(4). 267–288.
Wolfram, Walt. 1969. *A Sociolinguistic Description of Detroit Negro Speech*. Center for Applied Linguistics.
World Bank. 2018. *The World Bank: Cabo Verde Education Project (P164294)*. Project Information Document/Integrated Safeguards Data Sheet (PID/ISDS). The World Bank.
Wratil, Melani. 2009. Null arguments and homonymy flight in the development of Creole pronoun systems. In Peter K. Austin, Oliver Bond, Monik Charette, David Nathan & Peter Sells (eds.), *Proceedings of the Conference on Language Documentation and Linguistic Theory 2*. London: SOAS
Wratil, Melani. 2011. Uncovered pro – On the development and identification of null subjects. In Melani Wratil & Peter Gallmann (eds.), *Null Pronouns*. Berlin/Boston: de Gruyter.
Xia, Ting. 2018. *Visual Processing of Chinese Characters by Foreign Language Learners*. Ottawa, Ontario: Carleton University. (Masters' thesis).
Yim, Dongsun, Yoonhee Yang & Shinyoung Kim. 2015. Domain-Specific Working Memory Performance in Children with and without Specific Language Impairment. *Communication Sciences & Disorders* 20(1). 13–23.
Zanuttini, Raffaella. 1997. *Negation and Clausal Structure: A Comparative Study of Romance Languages*. Oxford University Press.
Zribi-Hertz, Anne. 1994. La syntaxe des clitiques nominatifs en français standard et en français avancé. *Travaux de Linguistique et de Philologie* 32. 131–148.

Zribi-Hertz, Anne & Lamine Diagne. 2002. Clitic Placement after Syntax: Evidence from Wolof Person and Locative Markers. *Natural Language & Linguistic Theory* 62.
Zwicky, Arnold M. & Geoffrey K. Pullum. 1983. Cliticization vs. Inflection: English N'T. *Language* 59(3). 502.